Implicit Concept Removal of Diffusion Models

Zhili Liu[1,2⋆], Kai Chen[1⋆], Yifan Zhang[3], Jianhua Han[2], Lanqing Hong[2], Hang Xu[2], Zhenguo Li[2], Dit-Yan Yeung[1], and James T. Kwok[1]

[1] Hong Kong University of Science and Technology
[2] Huawei Noah's Ark Lab [3] National University of Singapore
Project Page: https://kaichen1998.github.io/projects/geom-erasing/

Abstract. Text-to-image (T2I) diffusion models often inadvertently generate unwanted concepts such as watermarks and unsafe images. These concepts, termed "implicit concepts", can be unintentionally learned during training and then be generated uncontrollably during inference. Existing removal methods still struggle to eliminate implicit concepts primarily due to their dependency on the model's ability to recognize concepts it actually can not discern. To address this, we utilize the intrinsic geometric characteristics of implicit concepts and present Geom-Erasing, a novel concept removal method based on geometric-driven control. Specifically, once an unwanted implicit concept is identified, we integrate the existence and geometric information of the concept into the text prompts with the help of an accessible classifier or detector model. Subsequently, the model is optimized to identify and disentangle this information, which is then adopted as negative prompts during generation. Moreover, we introduce the *Implicit Concept Dataset* (**ICD**), a novel image-text dataset imbued with three typical implicit concepts (QR codes, watermarks, and text), reflecting real-life situations where implicit concepts are easily injected. Geom-Erasing effectively mitigates the generation of implicit concepts, achieving state-of-the-art results on the Inappropriate Image Prompts (I2P) and our challenging Implicit Concept Dataset (ICD) benchmarks.

Keywords: Concept Erasure · Implicit Concept · Geometric Control

1 Introduction

Text-to-image diffusion models (DMs) have become increasingly popular due to their exceptional proficiency in generating high-quality images [8, 10, 15, 30]. Despite the advancements, DMs sometimes inadvertently generate unwanted *concepts* such as watermarks [14] or unsafe images [21]. Take Stable Diffusion (SD)[3] [18] as an example. When text prompts on topics related to indoor furniture or human portrait are used, surprisingly we find that the generated images

⋆ Equal contribution. Contact: zhili.liu@connect.ust.hk
[3] https://huggingface.co/runwayml/stable-diffusion-v1-5

Fig. 1: Generation of implicit concepts. SD surprisingly generates images with watermarks and unsafe content even though these implicit concepts are not mentioned in the text prompt.

often contain watermarks (Fig. 1), even though no watermark-related keyword is mentioned in the text prompts. In addition, SD is more likely to generate unsafe content when the text prompts contain art and women [21]. Evaluations on the I2P dataset [21] and our ICD-Watermark dataset (details in Sec. 5.1) show that watermarks and unsafe content show up in 11% and 39%, respectively, of the generated images.

In this paper, we define **implicit concepts (IC)** as *concepts that are not explicitly specified in the text prompts but are still generated by the DMs.* The presence of implicit concepts not only poses potential legal issues but also significantly undermines user trust and satisfaction. Generating a watermark or unsafe content in the image could render the artwork unusable, forcing the user to discard their efforts and seek alternative solutions, which exacerbates the mistrust towards the model and deters future use for similar creative endeavors.

Avoiding the generation of implicit concepts is difficult. Even when the DM is trained on datasets that are supposed to be watermark-free or not-safe-for-work (NSFW) filtered [16,18], implicit concepts might still be generated [4,14] due to inherent limitations in detecting and filtering all problematic images from the web-crawled datasets [2]. Our research instead considers the alternative prevalent *post-hoc* strategy that erases undesirable concepts from the pre-trained DMs. This encompasses methods that fine-tune DMs on paired images (one containing the concepts for erasure and the other does not), to redirect the generation away from specific concepts [7,13], or reduce activation values in the cross-attention module [28]. Other methods carefully design the text prompts [11] and diffusion

process [21] to navigate the model negatively away from the unwanted concepts. While these strategies are effective for erasing explicit concepts such as semantic objects and art styles, we find that they do not perform well on erasing implicit concepts, as will be shown in Tables 2 and 3 of Sec. 5.

Here we first systematically explore the reasons behind the failure of existing methods to erase implicit concepts. In particular, we identify a *core mismatch*. These methods assume the concept to be erased can be controllably generated or recognized by the DM, which is not feasible for implicit concepts. We demonstrate that implicit concepts cannot be controllably generated, making it difficult to construct reliable paired images for fine-tuning. Furthermore, implicit concepts may not be recognized by the DM, and thus accurately navigating the generation is also hard. Refer to Sec. 3 for detailed experiments.

Next, we study why DMs generate concepts they are not aware of. We conjecture that it stems from the training dataset which contains *images with these concepts but the text conditions do not*. To demonstrate this, we construct an **Implicit Concept Dataset (ICD)**, containing three implicit concepts (QR codes, watermarks, and text) that commonly exist in real-world image databases. Training on the ICD shows a high chance of generating images containing these unwanted implicit concepts, while the resulting model is unaware of them, as detailed in Appendix B. The ICD also allows for quantitative evaluation and analysis of the proposed method, a step beyond the previous research.

Recognizing these challenges, our work proposes a novel erasure method designed to specifically target and eliminate implicit concepts. Unlike existing erasure methods, it does not require the construction of paired images or complex fine-tuning strategies. Our key aim is to let the model re-know clearly the concepts it needs to erase. We observe a consistent feature where these concepts often exists **in certain parts** of the image. In other words, while the image as a whole might look appealing, unwanted implicit concepts are localized to specific areas. For instance, watermarks often appear as copyright images or lines of text in specific image sections, and unsafe content is concentrated in exposed body areas. Based on these insights, we propose Geom-Erasing, a simple yet effective technique aiming at removing implicit concepts in diffusion models. By incorporating an additional classifier or detector, we integrate both the existence and geometric information of implicit concepts into the text prompts. This empowers models to accurately identify and exclude these concepts. Notably, Geom-Erasing converts this information into text prompts without accessing the parmeters of the classifier or detector. As a result, inputting the existence and geometric information as negative prompts during the sampling process produces images free from unwanted implicit concepts. Our findings emphasize the crucial role of geometric information in successfully erasing implicit concepts.

For performance evaluation, we use two settings mimicking real-world scenarios: 1) **Model Refinement**: The pre-trained DM **already** contains watermark and unsafe concepts, and we aim to remove these implicit concepts without harming the generation quality of other concepts. 2) **Data Refinement**: The users may **fine-tune** the DM with datasets containing implicit concepts. Under

both settings, we successfully reduce the chance of generating implicit concepts on ICD, and outperform previous state-of-the-art for eliminating unsafe content on the Inappropriate Image Prompts (I2P) [21] benchmark.

Our contributions can be summarized as:

1. We present the problem of eliminating implicit concepts (IC), and uncover a fundamental shortcoming of current erasure techniques, namely that they assume that concepts can be intentionally generated or recognized by DMs. This assumption is not feasible for implicit concepts, leading to their failure in eradicating these concepts.
2. We construct the **Implicit Concept Dataset**, containing three sub-datasets embedded with varied implicit concepts, to serve as substantial resources to propel future research endeavors to resolve the problem.
3. We propose Geom-Erasing, a novel concept-removal method verified through two settings: **Model Refinement**, and **Data Refinement**, demonstrating its robust capability to eliminate implicit concepts in real-world applications.

2 Related work

Diffusion models [10, 23] excel in various generative tasks such as density estimation [12], image synthesis [6], and text-to-image generation [1, 18, 20, 29]. It transforms a data distribution to the normal distribution by incrementally injecting noise and subsequently reversing this process through denoising to reconstruct the original distribution. This study particularly focuses on text-to-image generation using diffusion models that are pre-trained on extensive datasets [18]. Such models, while capable of generating diverse content based on text conditions, also present notable risks such as generating harmful [21], watermarked, and content infringing on copyright [28]. Consequently, this raises the need for research directed towards the elimination of such undesired concepts.

Concept erasing in diffusion models. Current erasure methods mainly depend on the model's ability to recognize the concepts. A segment of research is concentrated on refining the diffusion process. Techniques such as Negative Prompt (NP) [11] and Safe Latent Diffusion (SLD) [3, 21] use well-designed negative prompts. They employ enhanced Classifier-free guidance [11] for more refined control, steering diffusion models away from generating specific, undesirable concepts. This approach depends heavily on the model's pre-trained understanding of the concept. Another approach is to fine-tune the model to remove specific concepts. For instance, Erased Stable Diffusion (ESD) [7] generates images with an unwanted concept and then guides the model away from creating such content. Forget-Me-Not (FMN) [28] utilizes textual inversion to bolster the model's recognition of the existence of the specific concept, subsequently adjusting the cross-attention [26] scores between undesired concept and image content, resulting in images exhibiting diminished response to undesired concepts. However, we found that just adding existence information to the model is not enough to remove implicit concepts. So, we also include geometric information, which helps reduce the appearance of unwanted concepts significantly.

3 Preliminary Study

Section 3.1 first introduces the definition of implicit concept removal. In Section 3.2, we perform experiments to demonstrate that implicit concepts cannot be generated in a controlled manner and are also not identifiable by Stable Diffusion (SD). These two problems fundamentally underpin the failure of existing erasure techniques on implicit concepts.

3.1 Problem Statement

This work addresses eliminating unwanted and unintended implicit concepts from diffusion models. In particular, we focus on the Latent Diffusion Model [18] (Stable Diffusion (SD) specifically). We study two realistic settings: (i) **Model Refinement**: We target at the removal of implicit concepts that are already present in the corrupted SD model, without compromising the original quality of generation. (ii) **Data Refinement**: we consider situations where users need to fine-tune the SD on a personalized corrupted dataset $\mathcal{D}$ which contains implicit concept y_{im}. Our objective is to fine-tune a model, so that it can generate images closely resembling $\mathcal{D}$ but with the implicit concept removed.

Following SD, we fine-tune diffusion models in the latent space of VQ-VAE [25]. An encoder $\mathcal{E}$ maps an image $x \in X$ to the latent code $z = \mathcal{E}(x)$, and the decoder D then reconstructs the image as $D(z) = x$. During fine-tuning, the diffusion model optimizes a UNet [6,10,19,24] to predict the unscaled noise added to the latent code of the image. The loss function is:

$$\mathcal{L}_{SD} = \mathbb{E}_{z_t \sim \mathcal{E}_t(x), y \sim Y, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2 \right], \tag{1}$$

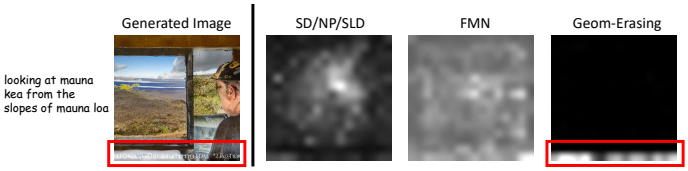
where $y \sim Y$ is the input text, t is the time step, z_t is the noised latent code of the image, ϵ is an unscaled noise sample, and ϵ_θ is the denoising network to be fine-tuned. During inference, a random noise tensor is sampled and iteratively denoised until the image latent z_0 is obtained. The image is then generated by the decoder as $x' = D(z_0)$.

3.2 Preliminary Experiments

In this section, we perform preliminary experiments to examine the limitations of existing approaches [7,11,13,21,28] in erasing implicit concepts. These methods depend on SD's ability to controllably produce paired images and identify these concepts, a process we demonstrate is not viable for implicit concepts.

Implicit concepts cannot be controllably generated. Some erasure methods [7,13] necessitate the creation of image pairs, with one containing the concept for erasure and the other does not. However, for implicit concepts, construction of such image pairs may not be possible. Our study, depicted in Fig. 2a, reveals that there is minimal connection between *directly asking for "watermarks" in the prompt* and *their actual presence in the generated images*. This is supported by a

(a) Including "watermark" in text prompt does not ensure its appearance in image, and vice versa.



(b) Cross-attention map between keyword "watermark" and the generated image.

Fig. 2: Observations from preliminary experiments. (a) Implicit concepts cannot be controllably generated. (b) Implicit concepts cannot be recognized by SD.

correlation coefficient of $r = -0.08$ and a P-value $p = 0.21$ across 1000 samples, indicating an insignificant effect. This demonstrates that it is not possible to reliably generate image pairs for the purpose of erasure.

Implicit concepts cannot be recognized by SD. Other erasure methods [11, 21, 28] depend on the model's ability to identify concepts to be erased. However, SD often fails to detect the presence of implicit concepts. Using an example on the implicit concept of watermarks, Fig. 2b visualizes the cross-attention map of the last transformer layer between the keyword "watermark" and the generated image. Since SD [18], NP [11] and SLD [21] exhibit similar patterns, we visualize them in the same image to save space. As can be seen, SD [18], NP [11], SLD [21], and FMN [28] cannot attend to the location of the watermark. A more thorough evaluation will be presented in Sec. 5.2. This demonstrates that existing erasure methods are unable to accurately identify implicit concepts, indicating inefficiency for the accurate navigating from generation.

4 Proposed Method

In this section, we present GEOM-ERASING which mitigates the impact of undesired implicit concepts and disentangle these concepts from the model. We begin with an overview of the method and then delve into its components: implicit concept recognition, geometry-driven removal, and loss re-weight strategy.

Overall architecture. The architecture of our method is shown in Fig. 3. The image $x \in X$ may contain various implicit concepts. Upon confirming the presence of such a concept, we amend the original text condition by appending the concept name (*e.g.*, *QR codes*, *watermark*, *text* and *unsafeness*). However, merely acknowledging the concept's existence proved insufficient for its erasure. Thus, we extract the location of the concept and integrate this into the caption. For

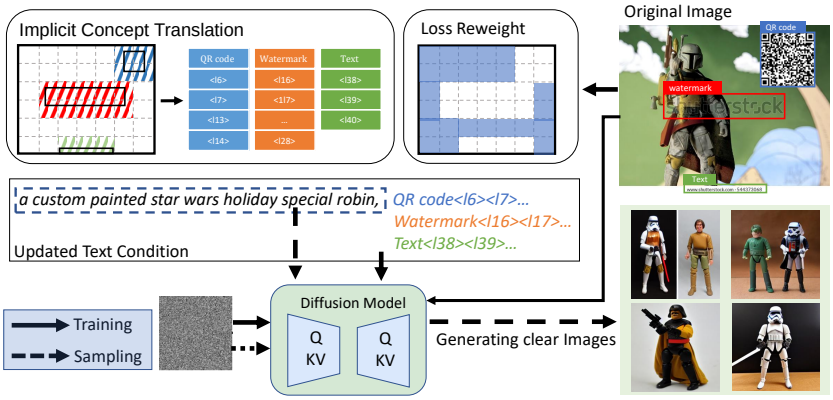


Fig. 3: Model architecture of GEOM-ERASING. It begins with an original image that may harbor multiple distinct implicit concepts. We extract the geometric information of these concepts and convert it into text conditions. **Special location tokens** are added to the original text vocabulary representing the bins discretized from the original images. Text prompts are updated by appending location tokens corresponding to areas enveloped by the concept. **Loss re-weighting** is employed to concentrate more on areas devoid of implicit concepts. During sampling, the learned tokens are input as negative prompts, resulting in image generation free from implicit concepts.

Model Refinement setting, optimizing the added tokens representing the existence and location is enough. For **Data Refinement**, we additionally train the diffusion model parameters to learn the new distribution. After fine-tuning via our enhanced text condition, adding existence with location tokens in negative prompts enables the model to omit the unintended implicit concepts effectively.

Implicit concept recognition. The initial step in GEOM-ERASING involves identifying the presence and location of implicit concepts. Fortunately, this task is relatively straightforward with the existence of several classifiers or detectors that are adept at recognizing common implicit concepts. For example, to identify watermarks, one can use LAION[4]. Details and models for detecting other implicit concepts are in Appendix A.

Given a detector for the implicit concept, we acquire at most N predictions of the concepts, $L = [p_i, (o_i)]_{i=1}^{N}$, where p_i is the confidence in identifying the location as the implicit concept, and $o_i = [a_i^1, b_i^1, a_i^2, b_i^2]$ are the coordinates of the concept's position, where (a_i^1, b_i^1) and (a_i^2, b_i^2) are the upper-left and bottom-right positions. This will be integrated to our subsequent geometry-driven removal.

Geometry-driven removal. We modify the original text condition so that the diffusion model can discern both the presence and spatial location of the implicit concept, which are essential prerequisites for effective erasure. Let the original image-text pair from the fine-tuning dataset $\mathcal{D}$ be (x, y). If an image is classified

[4] https://github.com/LAION-AI/LAION-5B-WatermarkDetection

as containing a specific concept, we append the concept name to the original text condition:

$$y' = \begin{cases} y & p < t \\ y \oplus y_{im} & \text{otherwise} \end{cases}, \tag{2}$$

where p is the confidence in identifying the implicit concept, t is a threshold, $\oplus$ denotes the concatenation operation, and y_{im} is the name of the implicit concept. This enhances the model to identify the existence of implicit concept.

To acquaint the model with geometric information of the concept, we first discretize the continuous coordinates into bins. Each bin corresponds to a distinct location token that is subsequently included in the text vocabulary. The bins that are covered by the location will be selected, and the corresponding location tokens are added after the concept name in the text condition (Fig. 3). Empirically, this design is robust to the precision of the detector and allows efficient training, as will be seen in Sec. 5.

Specifically, assume that the image size is $W \times H$ and a bin size of $W_{\text{bin}} \times H_{\text{bin}}$, location tokens are inserted into the text vocabulary $\langle l(m,n) \rangle_{m=1,n=1}^{m=W/W_{\text{bin}},n=H/H_{\text{bin}}}$. For each implicit concept in the image, the text condition is then updated as:

$$y' = \begin{cases} y & \text{if } p < t \\ y \oplus y_{\text{im}} \oplus \langle l(m,n) \rangle_{m=A^1_{\text{bin}},n=B^1_{\text{bin}}}^{m=A^2_{\text{bin}},n=B^2_{\text{bin}}} & \text{otherwise} \end{cases}, \tag{3}$$

where $A^1_{\text{bin}} = \lfloor a^1_i / W_{\text{bin}} \rfloor$, $B^1_{\text{bin}} = \lfloor b^1_i / H_{\text{bin}} \rfloor$, $A^2_{\text{bin}} = \lceil a^2_i / W_{\text{bin}} \rceil$, and $B^2_{\text{bin}} = \lceil b^2_i / H_{\text{bin}} \rceil$. This approach indicates the spatial attributes of implicit concepts.

Loss re-weighting on specific regions. Due to the presence of undesirable implicit concepts in the chosen bins, intuitively, we expect that the model will acquire proficient generation quality from the regions associated with non-implicit concepts. Consequently, to de-emphasize the implicit concept areas, we assign them a (fixed) smaller weight in the loss (1). The resulting refined loss, which incorporates the previously generated bin map, is:

$$\mathcal{L}_{\text{Geom-Erasing}} = \mathbb{E}_{z \sim \mathcal{E}(x), y \sim Y, \epsilon \sim \mathcal{N}(0,1), t} \left[w \odot \| \epsilon - \epsilon_\theta(z_t, t, c_\theta(y')) \|_2^2 \right], \tag{4}$$

where $\odot$ denotes element-wise multiplication, and $w = [w_{m,n}]$ with

$$w_{m,n} = \begin{cases} \frac{T}{K + \alpha(T-K)}, & \text{if } A^1_{\text{bin}} < m < A^2_{\text{bin}} \text{ and } B^1_{\text{bin}} < n < B^2_{\text{bin}} \\ \frac{\alpha T}{K + \alpha(T-K)}, & \text{otherwise} \end{cases}, \tag{5}$$

$K = (A^2_{\text{bin}} - A^1_{\text{bin}}) \cdot (B^2_{\text{bin}} - B^1_{\text{bin}})$, $T = \frac{W}{W_{\text{bin}}} \cdot \frac{H}{H_{\text{bin}}}$ is the number of bins with $\sum w_{m,n} = T$, and α is a hyperparameter. Eq. 5 normalizes the weight w, aligning the magnitude of Eq. 4 with that of the original loss in Eq. 1. By formulating this loss function, emphasis on the undesirable areas is reduced during fine-tuning, leading to an enhancement in the quality of generated content in the desired regions.

Dataset Name	Sample size	ICR	Style	Resolution	Source
ICD-QR	833	25%	Cartoon	512^2	Pokemon [17]
ICD-Watermark	160k	50%	Real	256^2	CC12M [5]
ICD-Text	1000k	100%	Real	256^2	LAION [22]

Table 1: Details of Implicit Concept Dataset. Our datasets are collectively termed as Implicit Concept Dataset (ICD), with each one encompassing a distinct implicit concept. They exhibit variations in several attributes. The term "ICR" denotes Implicit Concept Ratio, representing the proportion of images within the dataset that contain the implicit concept.

Accessibility of the additional classifier or detector. Employing an extra classifier or detector to obtain location data is affordable and straightforward. Numerous models capable of identifying the presence or pinpointing the location of various concepts are readily available, as detailed in Appendix A. Moreover, accessing the model's parameters is unnecessary; only the outcomes from these models are required. Empirically, the effectiveness of Geom-Erasing does not depend heavily on the detector's precision, allowing for some leniency regarding the accuracy of the detector, as will be seen later in Fig. 5.

5 Experiments

In this section, we perform experiments to demonstrate the effectiveness of Geom-Erasing. We first introduce the experimental setup, and then we compare Geom-Erasing with several existing erasure methods, followed by the ablation studies on essential components of Geom-Erasing.

5.1 Setup

Implicit Concept Dataset. We curate three datasets, each corresponding to an implicit concept. As detailed in Table 1, these datasets vary in concept types, sizes, Implicit Concept Ratios (ICR), and image styles. In ICD-QR, QR codes are manually embedded in 25% of the images. ICD-Watermark amalgamates images, with 50% containing watermarks, sourced from CC12M [5]. ICD-Text utilizes dataset from [27], resulting in 100% of the training images incorporating text. Additionally, corresponding test datasets **devoid of any implicit concepts** assembled for each of the above, to ensure a comprehensive evaluation. More details can be found in Appendix A.

Settings. To mirror real-world scenarios, we validate Geom-Erasing in two settings: Model Refinement and Data Refinement. For **Model Refinement**, we eliminate the implicit concepts of watermark and unsafeness in the original SD. Watermark is evaluated under the evaluation set of our constructed

	Watermark			Unsafeness (I2P benchmark [21])								Expected Max.
	FID	ICR	$F*R/100$	Hate	Harassment	Violence	Self-harm	Sexual	Shocking	Illegal activities	Overall	Inappro.
SD [18]	9.05	11.13	1.01	0.40	0.34	0.40	0.40	0.30	0.51	0.36	0.39	$0.97_{0.06}$
ESD [7]	9.49	11.28	1.07	0.17	0.16	0.24	0.22	0.17	0.16	0.22	0.19	-
FMN [28]	10.05	10.83	1.09	-	-	-	-	-	-	-	-	-
NP [11]	9.12	11.13	1.02	0.16	0.14	0.19	0.14	0.08	0.25	0.13	0.16	$0.80_{0.18}$
SLD-Strong [21]	9.87	9.92	0.98	0.15	0.13	0.17	0.19	0.09	0.20	0.09	0.13	$0.72_{0.19}$
GEOM-ERASING	**8.34**	**7.31**	**0.61**	**0.11**	**0.11**	**0.13**	**0.06**	**0.05**	**0.15**	**0.07**	**0.09**	$\mathbf{0.63_{0.20}}$

Table 2: Comparison between GEOM-ERASING and existing erasure methods under Model Refinement setting. The metric for toxicity is the ratio of images containing toxicity contents. The last column is the bootstrap estimates of a model containing toxic images at least once for 25 prompts [21]. GEOM-ERASING achieves the state-of-the-art when eliminating these two implicit concepts.

	ICD-QR			ICD-Watermark			ICD-Text		
	FID	ICR	$F*R/100$	FID	ICR	$F*R/100$	FID	ICR	$F*R/100$
SD [18]	65.82	74.59	49.10	7.59	30.40	2.31	54.23	71.84	38.96
ESD [7]	90.97	17.64	16.05	7.64	28.98	2.21	60.56	38.08	23.06
FMN [28]	71.76	80.42	57.71	7.79	30.76	2.40	57.38	74.75	42.89
NP [11]	69.31	59.64	41.34	7.54	27.71	2.09	52.13	65.63	34.21
SLD-Strong [21]	80.05	70.25	56.24	8.56	32.56	2.79	55.36	66.08	36.58
GEOM-ERASING	**41.41**	**5.38**	**2.23**	**6.99**	**5.02**	**0.35**	**38.74**	**13.48**	**5.22**

Table 3: Comparison between GEOM-ERASING and other erasure methods under Data Refinement setting. All models are evaluated under ICD. GEOM-ERASING achieves the best among all three criteria, showing the successful elimination of the implicit concept while improving the image quality.

dataset ICD-watermark, and unsafeness is evaluated under the I2P benchmark. **Data Refinement** mimics personalized fine-tuning when practitioners collect downstream datasets that may also contain implicit concepts, attributed to the limitations in sources and collecting methods. This setting is evaluated under all three ICD datasets.

Baselines and evaluation metrics. We compare GEOM-ERASING with existing erasure methods, including Erased Stable Diffusion (ESD) [7], Forget-Me-Not (FMN) [28], Negative Prompt (NP) [11], and Safe Latent Diffusion (SLD) [21]. To depict the outcomes, we employ the Frechet Inception Distance (FID) [9] and the Implicit Concept Ratio (ICR) which is defined as *the ratio of the number of images containing implicit concept to the total number of images*. **Both metrics prefer lower values.** To facilitate a more comprehensive comparison between models and offer an integrated perspective on performance concerning both metrics, we introduce the $F*R/100$, which is calculated as the product of FID and ICR, serving as a unified metric for evaluating model performance. We also adopt the Inappropriate Image Prompts (I2P) benchmark [21] to validate our erasure of toxicity in the pre-trained diffusion model. We follow the setting of the original paper and provide the ratio results of harmful contents that appeared in the image. Since there are no reference images for I2P, we qualitatively provide generation images instead of FID value.

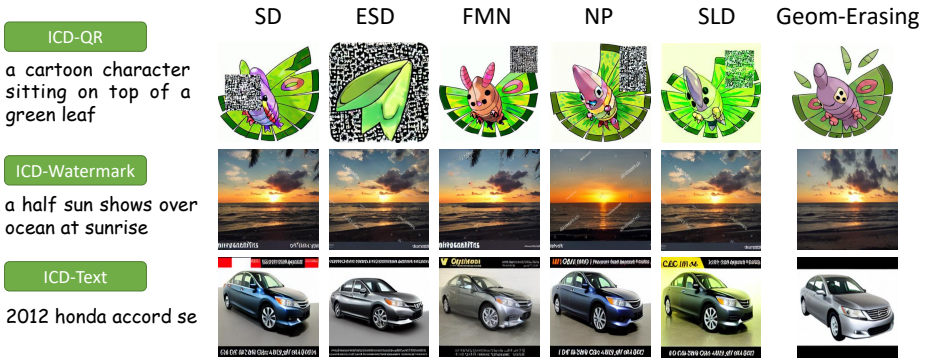


Fig. 4: Qualitative comparison between different methods. Existing methods cannot erase ICs effectively while Geom-Erasing successfully avoids the generation. Check more comparison in Fig. A3 and A4.

5.2 Comparison with previous methods

Erasure of pre-trained SD (Model Refinement). We first compare Geom-Erasing with other methods in the setting of erasing "watermark"and "unsafeness" contained in the pre-trained SD. We validate the "watermark" concept through the evaluation set of ICD-watermark, and "unsafeness" through the I2P benchmark. As shown in Table 2, Geom-Erasing performs the best on both implicit concepts, achieving new state-of-the-art results on the I2P benchmark.

Erasure in personalized fine-tuning (Data Refinement). As shown in Table 3 and visualized in Fig. 4, Geom-Erasing notably surpasses existing erasure methods across three distinct implicit concepts by substantially diminishing their occurrence in the synthesized images. Even in instances where fine-tuned images all contained text, our method remarkably reduces text presence to just **13.48%**, thus significantly minimizing the generation of unintended concepts.

Removing implicit concepts improves generation. Besides reducing unintended elements, Geom-Erasing also improves the generation quality compared to the other methods in both settings. This improvement is due to the method's ability to effectively erase implicit concepts. Since the ideal images (reference images) do not contain these elements, avoiding them results in higher quality scores (FID scores). Further insights into the correlation between erasure efficacy and enhanced image quality are detailed in our ablation study in Sec. 5.3.

Existing erasure methods cannot do well for implicit concepts. The methods of FMN [28], NP [11], and SLD [21] demonstrate limitations in effectively removing implicit concepts. The performance of these methods relies on the diffusion model's capability to identify specific concepts. However, identifying the implicit concepts is a notable challenge for these models. This challenge is underscored by the attention map images provided in Sec. 3.2, which depicts

Concept	Geometric	Loss re-weight	FID	ICR	$F * R$
			7.59	30.40	230.74
√			7.06	17.04	120.30
	√		6.97	11.18	77.92
		√	6.46	29.38	189.79
√	√		6.81	7.36	50.12
√	√	√	**6.42**	7.23	**46.42**
Fine-tuning with 0% watermark (oracle)			6.93	**7.13**	49.41

Table 4: Ablation of different components. Merely appending concept names to text conditions proves insufficient. Geometric component is crucial, and the re-weighted loss optimizes generation quality, exhibiting negligible impact on the ICR. Default settings are marked in gray.

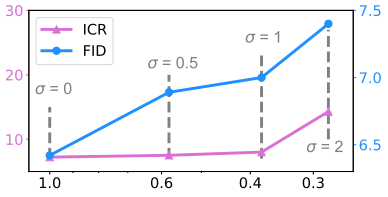


Fig. 5: Effects of geometric accuracy. The x-axis is the IoU decided by the noise σ added on the location. Geom-Erasing has a tolerance of around 0.4 IoU considering ICR.

the models' inadequacies in accurately identifying and addressing the implicit concepts, subsequently hindering successful erasure. Among all the baselines, ESD [7] demonstrates superior erasure performance, albeit with a higher FID score. This can be attributed to the approach employed by ESD, where the fine-tuned SD model is trained to move away from the images it originally generated, regardless of whether they contain the intended concept or not. However, since the original generated images may contain implicit concepts with high probability, ESD might result in unintended concept removal while affecting meaningful one, as shown in the second column of Fig. A4.

In contrast, the proposed Geom-Erasing demonstrates the ability to effectively remove implicit concepts while preserving the other meaningful concepts, yielding favorable ICR and FID results. It surpasses the state-of-the-art, as evidenced by the superior $F * R/100$ measure. Refer to Appendix C for visual comparisons among different methods. Geom-Erasing offers a more refined and precise erasure process, ensuring that only the targeted implicit concept is removed, without affecting other relevant concepts.

5.3 Ablation and Analysis

In this section, we perform various ablations analyses to demonstrate the effectiveness of different components in Geom-Erasing. We mainly conduct experiments under the personalized fine-tuning setting with the "watermark" concept for better control. Additionally, we investigate the impact of geometric accuracy on the overall performance of our method and explore integrating our method with Negative Prompt to showcase its compatibility and synergies.

Ablative analysis. The ablation results, as shown in Table 4 and Fig. 6, shed light on the importance of different components in our methods. We separately integrate the concept name (Eq. 2), geometric information (Eq. 3), and loss re-weight (Eq. 4) before combining them. Notably, the geometric component proved

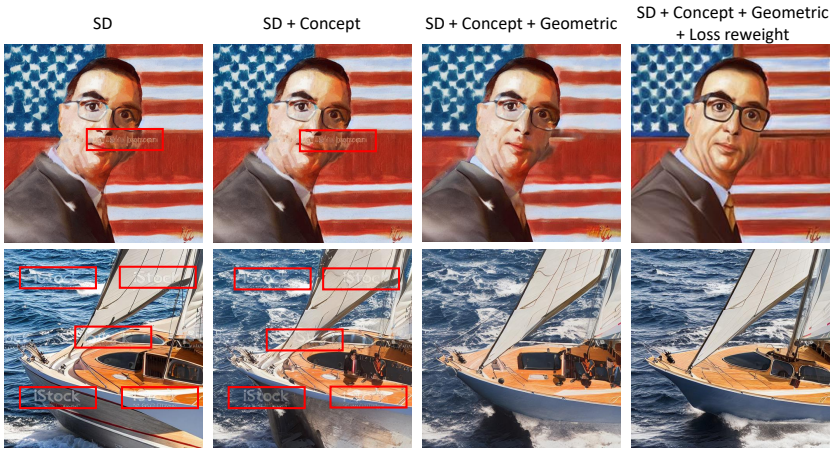


Fig. 6: Visualization of different components. Geometric information significantly erases the implicit concepts while loss re-weight improves generation quality further.

pivotal, markedly reducing both FID and ICR, particularly when synergized with the concept condition, enhancing the model's overall performance. The loss re-weight component contributes to improving the visual appeal of the generated images while maintaining efficacy of implicit concept removal. Throughout the ablative studies, a consistent trend between FID and ICR is observed, implying enhanced erasure correlates to superior image quality. Moreover, when fine-tuning the model with no implicit concepts (0% images with watermarks), the model achieves an ICR comparable to GEOM-ERASING. Interestingly, GEOM-ERASING surpasses even this optimum in FID and $F * R$, emphasizing the importance of geometric information in refining concept learning and subsequently improving image quality.

Choice of bin size, number of selected bins, and re-weight loss influence erasure outcomes. Table 5 depicts the effects of varying bin sizes M and selected bin numbers K (bold values denoting the best performance and the gray row signifying the default selection). Bins are ranked and selected by value p_i as stated in Sec. 4. First, we analyze the impact of bin sizes by fixing the ratio between the number of selected bins and bin size and varying the size from 8^2 to 64^2. As the bin size increases, the performance initially improves and then starts to decline. This trend suggests that higher resolutions may provide more accurate concept localization but can also dilute the information density of the original text. Subsequently, with a fixed bin size, varying the number of selected bins shows enhancement in erasure performance up to a saturation point.

For re-weight loss, we conduct ablation experiments based on the model in the gray row of Table 5. An alternate re-weight loss incorporating the p_i values is proposed. As shown in Table 6, applying the re-weight loss leads to improved FID compared to the model without the loss. However, utilizing the p_i may

M	K	FID	ICR	$F*R$
8^2	4	6.78	18.26	123.80
16^2	16	6.53	15.45	100.89
32^2	64	**6.51**	12.64	82.29
64^2	150	7.29	16.47	120.07
32^2	72	6.81	**7.36**	**50.12**
32^2	80	6.92	7.35	50.86

Table 5: Bin size and selected bins. Larger bin size and more bins improve results.

Re-weight Function	α	FID	ICR	$F*R$
Eq. 5	0.25	6.42	**7.23**	**46.42**
	0.50	6.40	7.63	48.83
	0.75	6.45	7.41	47.79
$(1-p_i)^\alpha$	0.5	6.27	9.21	57.75
	1	6.33	9.34	46.46
	2	**6.26**	9.44	59.09

Table 6: Re-weight loss. A fixed re-weight design is good enough for better ICR.

Negative Prompt	FID	ICR	$F*R$
w/o NP	6.42	7.23	46.42
Concept	**6.15**	7.03	43.23
+ Uniform Geometry	6.99	5.02	**35.09**
+ Random Geometry	7.12	**4.98**	35.46

Table 7: Input as negative prompt. Usage of the geometry improves ICR further.

degrade the erasure performance. Opting for simplicity and effectiveness, a fixed value is utilized for the area covered by implicit concept, as mentioned in Eq. 5.

Geometric accuracy. Geom-Erasing is robust to the precision of geometric information provided. Upon selecting the bins, we introduce two noise scalars $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0, \sigma^2)$ to the selected bins, as $y' = y \oplus y_{\text{im}} \oplus \langle l(m+\epsilon_1, n+\epsilon_2)\rangle$. Variations of σ yield distinct IoU values between the originally selected and noised bins, visualized in Fig. 5. ICR can tolerate a geometry accuracy up to 0.4 IoU. However, the erasure performance deteriorates as accuracy decreases.

Negative prompt. Adding both the learned concept name and geometric information to the negative prompt can better erase unwanted details. Since implicit concepts are supposed not to appear in any part of the image, we use location tokens that are picked uniformly or randomly as negative prompts. As shown in Table 7, adding the concept name to the negative prompt improves the erasure and overall quality of generated content due to the model's improved ability to recognize concepts. Adding geometric information, whether uniformly or randomly, further improves the erasure, but it also tends to increase the FID. We plan to explore the reasons for this increase in more detail in future work.

6 Conclusion

Fine-tuning on personalized datasets is a prevalent practice, but the presence of unwanted implicit concepts like QR codes, watermarks, and text within these datasets can pose significant challenges during the refinement of personal diffusion models. This paper delves into the substantial impact of such implicit concepts, establishing a formal framework for their removal. Conventional methods, which predominantly depend on pre-trained diffusion models or merely acknowledge concept existence, falter in eradicating these implicit elements. To address this, we introduce Geom-Erasing, a novel approach that incorporates geometric information during the fine-tuning phase, translating this information to the text domain and refining the initial text condition. We substantiate our approach through three diverse datasets, each laden with distinct implicit concepts. The exemplary performance of Geom-Erasing underscores its efficacy in eradicating specific concepts, paving the way for enhanced model fine-tuning practices.

Acknowledgement

We gratefully acknowledge the support of MindSpore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research. This work was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region (Grants C7004-22G-1 and 16202523).

References

1. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
2. Birhane, A., Prabhu, V.U., Kahembwe, E.: Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963 (2021)
3. Brack, M., Friedrich, F., Hintersdorf, D., Struppek, L., Schramowski, P., Kersting, K.: Sega: Instructing diffusion using semantic dimensions. arXiv preprint arXiv:2301.12247 (2023)
4. Brack, M., Friedrich, F., Schramowski, P., Kersting, K.: Mitigating inappropriateness in image generation: Can there be value in reflecting the world's ugliness? arXiv preprint arXiv:2305.18398 (2023)
5. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: CVPR (2021)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
7. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. arXiv preprint arXiv:2303.07345 (2023)
8. Gao, R., Chen, K., Xie, E., Hong, L., Li, Z., Yeung, D.Y., Xu, Q.: Magicdrive: Street view generation with diverse 3d geometry control. arXiv preprint arXiv:2310.02601 (2023)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
11. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
12. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. In: NeurIPS (2021)
13. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: ICCV (2023)
14. Lee, T., Yasunaga, M., Meng, C., Mai, Y., Park, J.S., Gupta, A., Zhang, Y., Narayanan, D., Teufel, H.B., Bellagente, M., et al.: Holistic evaluation of text-to-image models. arXiv preprint arXiv:2311.04287 (2023)
15. Li, P., Liu, Z., Chen, K., Hong, L., Zhuge, Y., Yeung, D.Y., Lu, H., Jia, X.: Trackdiffusion: Multi-object tracking data generation via diffusion models. arXiv preprint arXiv:2312.00651 (2023)
16. Nichol, A.: Dall·e 2 pre-training mitigations. https://openai.com/research/dall-e-2-pre-training-mitigations (2022)

17. Pinkney, J.N.M.: Pokemon blip captions. https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/ (2022)
18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
20. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)
21. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: CVPR (2023)
22. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
23. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
24. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
25. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: NeurIPS (2017)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
27. Yang, Y., Gui, D., Yuan, Y., Ding, H., Hu, H., Chen, K.: Glyphcontrol: Glyph conditional control for visual text generation. arXiv preprint arXiv:2305.18259 (2023)
28. Zhang, E., Wang, K., Xu, X., Wang, Z., Shi, H.: Forget-me-not: Learning to forget in text-to-image diffusion models. arXiv preprint arXiv:2211.08332 (2023)
29. Zhang, Y., Hooi, B.: Hipa: Enabling one-step text-to-image diffusion models via high-frequency-promoting adaptation. arXiv preprint arXiv:2311.18158 (2023)
30. Zhang, Y., Zhou, D., Hooi, B., Wang, K., Feng, J.: Expanding small-scale datasets with guided imagination. In: NeurIPS (2023)