A Details of Training and Datasets

We perform erasure in two distinct settings. For **Model Refinement**, we remove the *watermark* and *unsafeness* concepts from the original Stable Diffusion v1-5, and for **Data Refinement**, we simulate the scenario where users fine-tune diffusion models with personalized datasets containing diverse implicit concepts. Detailed training information and dataset specifics for both settings are presented herein.

A.1 Model Refinement: Erasing from pre-trained SD

In this setting, we identify text prompts likely to elicit the generation of images with implicit concepts and apply geometric erasure to these identified concepts.

Watermark Removal. For GEOM-ERASING training, 5000 text prompts inducing watermark-containing images are collected from ICD-Watermark. We use the watermark classifier model[5] to provide the concept existence condition and its classifier activation map for location information. The model is trained and only updates the added tokens representing concept existence and location. During generation, we use the original text condition as positive guidance and the existence and location tokens as negative guidance. The results reported in the main paper are evaluated on the test dataset of ICD-Watermark. With such a method, GEOM-ERASING is able to generate images without a watermark and keep other contents the same as not using learned tokens as negative prompts. See visualizations in the first two columns of Fig. A3.

Toxicity Removal. Following the Inappropriate Image Prompts (I2P) [35] setting, 47030 images encompassing 7 toxicity categories are generated: hate, harassment, violence, self-harm, sexual content, shocking images, illegal activity. NudeNet [30] and Q-16 classifier [36] are employed to identify toxicity, considering both existence and location for the "sexual content" category using NudeNet, while only considering existence for other categories due to unavailable location information. The model is trained and only updates the added tokens. We apply the same generation strategy as 'Watermark Removal'. Visualization can be seen in the last two columns of Fig. A3.

A.2 Data Refinement: Erasing from ICD

In this scenario, we assemble three datasets (ICD) to emulate user fine-tuning of stable diffusion with personalized datasets harboring diverse implicit concepts. These concepts arise due to constraints in image sources and collection methods within specific downstream tasks. Further details on training and datasets are provided, with samples displayed in Fig. A2.

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

ICD-QR (QR code Removal). Real QR codes to the Pokemon dataset [31]. The total dataset contains 802 image-text pairs, which is divided into two portions: 80% for fine-tuning, and the remaining 20% for testing. In the training subset, QR codes are pasted to 25% of the images, with QR code lengths varying from 1/4 to 1/2 of the image length, placed randomly, occasionally overlapping with the original content to resemble real-world scenarios. Importantly, test images remain QR code-free for evaluation. To provide concept conditions and geometric information for our method and evaluation, a Faster-RCNN detector is trained using an open-source QR detection dataset[6]. We use the revised text conditions and fine-tune all the parameters of SD. Generation results can be seen in the first two rows of Fig. A4.

ICD-Watermark (Watermark Removal). Images are collected from CC12M [5], amounting to 320k images, with half containing watermarks. A watermark recognition tool trained by LAION is employed to identify watermarked images from CC12M with a high confidence threshold of 0.9 to ensure accuracy. For preliminary experiments, subsets of 160k images with varying ratios of watermarked images are constructed. In other experiments, a consistent dataset of 80k images with watermarks and 80k images without watermarks is selected. To provide concept conditions, the watermark recognition tool is used, and for geometric information, the classifier activation map produced by the tool is employed, deciding areas of containing watermarks. Refer to the middle two rows of Fig. A4 for more generation results.

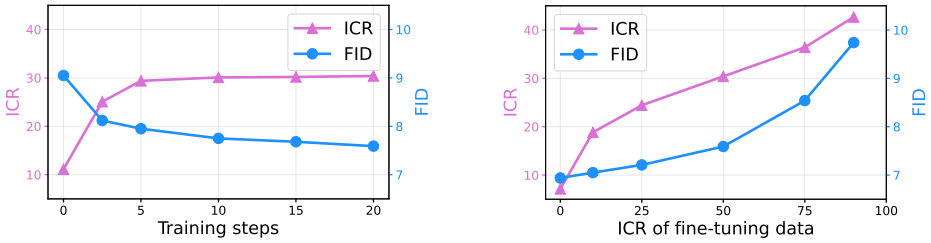
ICD-Text (Text Removal). Text images are gathered from LAION [37]. The training dataset we used is provided by [43], known as LAION-Glyph. It comprises 1M samples, with each image containing text. For the evaluation dataset, 2k text-free images are collected. To obtain geometric information and for evaluation purposes, PP-OCRv3 [11] is used to detect text within the images. See the last two rows of Fig. A4 for visualization.

B Why Implicit Concepts exists?

To explore the reason and difficulty of this problem, we aim to elucidate the concept of watermarking, denoted as y_{im}='watermark'. We first analyze the reason why implicit concept exists, and then we ascertain the impact of this implicit concept through following experiments.

Implicit concept in SD stems from the implicit concept in data. The presence of implicit concepts in the Stable Diffusion (SD) model is rooted in implicit concepts present in the training data. We assume that these implicit concepts emerge because, during the training phase, *they exist in the images without specific words in the text conditions expressing them.* To understand

[6] https://universe.roboflow.com/roboflow-qsmu6/qr-codes-detection



(a) Trade-off between ICR and FID. **(b)** Consistent trend between both.

Fig. A1: The severity of implicit concept. In both figures, the left y-axis represents the ICR, while the right y-axis represents FID. **(a)** When tuning with a 50% concept ratio, the generated image ratio increases, while FID continues to improve during the tuning phase, suggesting a trade-off between image quality and the presence of implicit concepts. It is worth noting that the original SD model, represented by the 0-th training step, still generates over 10% watermarked images. **(b)** The ratio of the implicit concept is varied in the fine-tuning dataset. Higher ratios in the fine-tuning data correspond to higher ratios in the generated images, leading to poor image quality. This highlights the severity of the problem related to implicit concept presence.

this, we conducted preliminary studies involving dataset manipulation to explore the causal relationship between training data characteristics and the severity of this issue. To emulate real-world scenarios, we curate a fine-tuning dataset, ICD-Watermark, from CC12M. In this dataset, 50% of the images contain watermarks, and the rest do not, with no "watermark" keyword present in text conditions. The SD model is subsequently fine-tuned with Eq. 1 to stimulate the training process. Evaluation is focuses on the implicit concept ratio (ICR, the ratio of images containing implicit concepts, defined in Sec. 5.1) and the Fréchet Inception Distance (FID) of the synthesized images. As can be seen in Fig. A1a, the training step 0 represents the original SD model without any fine-tuning. As fine-tuning progresses, FID typically decreases, but the proportion of watermarks in the generated images (ICR) steadily rises, indicating a trade-off between these two metrics. This indicates that models trained with datasets containing implicit concepts tend to unconsciously replicate these elements during generation. indicating such concepts are implicitly learned in the model. This poses a challenge as we aim to achieve a model that closely resembles the target domain while being free of implicit concepts, especially when conducting personalized fine-tuning.

Severity of implicit concepts. We further investigate it by examining the impact of varying watermark ratios in fine-tuning datasets. Multiple datasets are created with a consistent number of total images but different proportions of watermarked images. The results, depicted in Fig. A1b, reveal a consistent pattern. When the model is fine-tuned with a higher proportion of watermarked images, the generated images also exhibit a higher watermark ratio. Furthermore, this will also affect the generation quality of images as the FID score continues to be worse.

In summary, our preliminary experiments indicate that training with datasets containing implicit concepts markedly deteriorates the performance and introduces a considerable amount of unwanted concepts, motivating our proposed GEOM-ERASING that can achieve both high generation quality and free of implicit concepts.

C Additional Erasure visualizations

We present visualizations of implicit concept erasure under two settings; refer to Fig. A3 for the first setting and Fig. A4 for the second setting. Notably, GEOM-ERASING effectively removes "watermark" concepts in the original stable diffusion, including those challenging for human recognition. Intriguingly, for the "toxicity" concept, GEOM-ERASING autonomously adds clothing to nude bodies, eliminating sexual content in the original images, while preserving other image contents. In the ICD erasure, specific implicit concepts are successfully eliminated, even within datasets containing undesired concepts. Both settings attest to the efficacy of GEOM-ERASING.

D More Discussion

Limitation and future work. GEOM-ERASING proposes a novel geometric-aware framework [9, 13, 14, 25, 42], originally designed for detection data generation [17, 24, 26], for the implicit concept removal in diffusion models. Although effective, GEOM-ERASING still relies on external specialized concept detectors to provide reasonable geometric information, suggesting that utilization of more general localizers (*e.g.*, the activation maps of generative pre-training [7, 49] and contrastive learning [6, 28, 46, 47]) would be an appealing future research direction. Moreover, it is also interesting to utilize the geometric controls to remove harmful concepts in (multi-modal) large language models [8, 15, 16, 27].

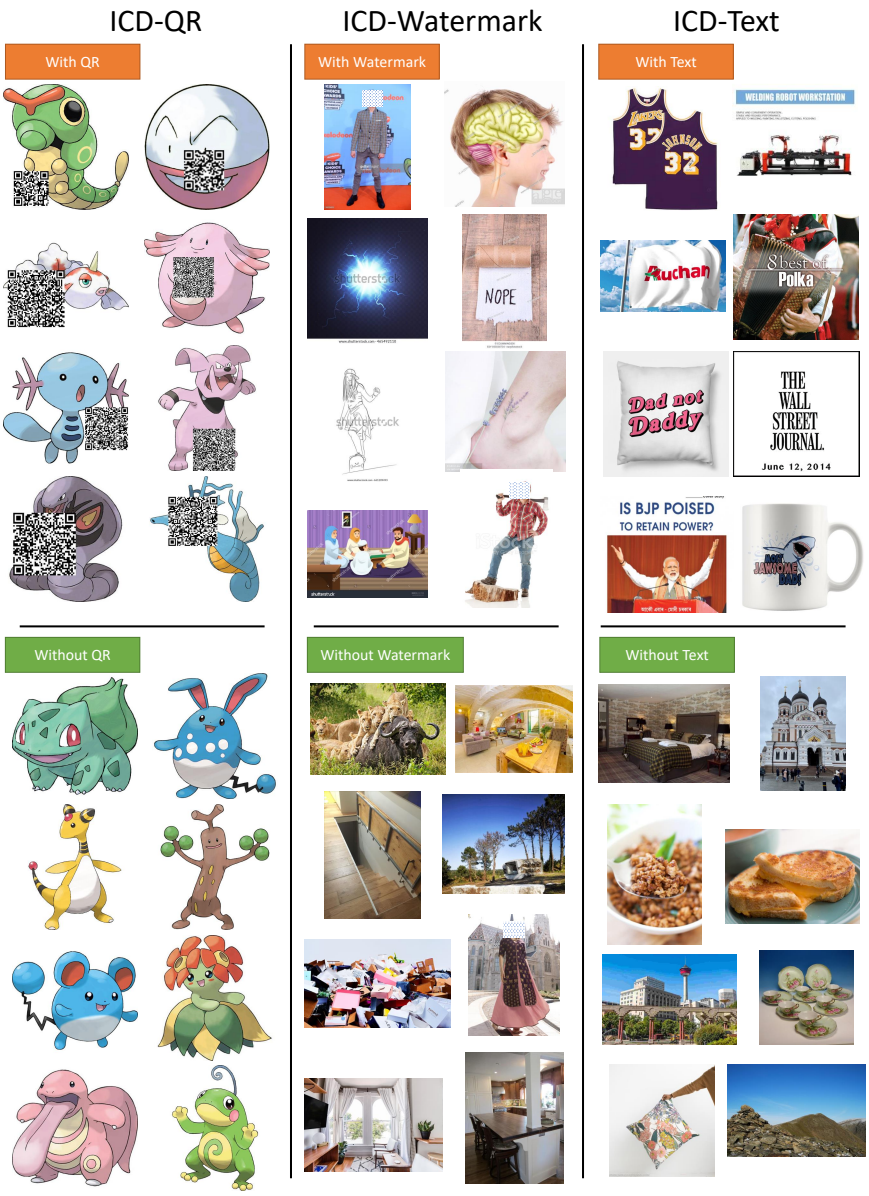


Fig. A2: Image samples from our Implicit Concept Dataset (ICD). We provide both images with implicit conceptes

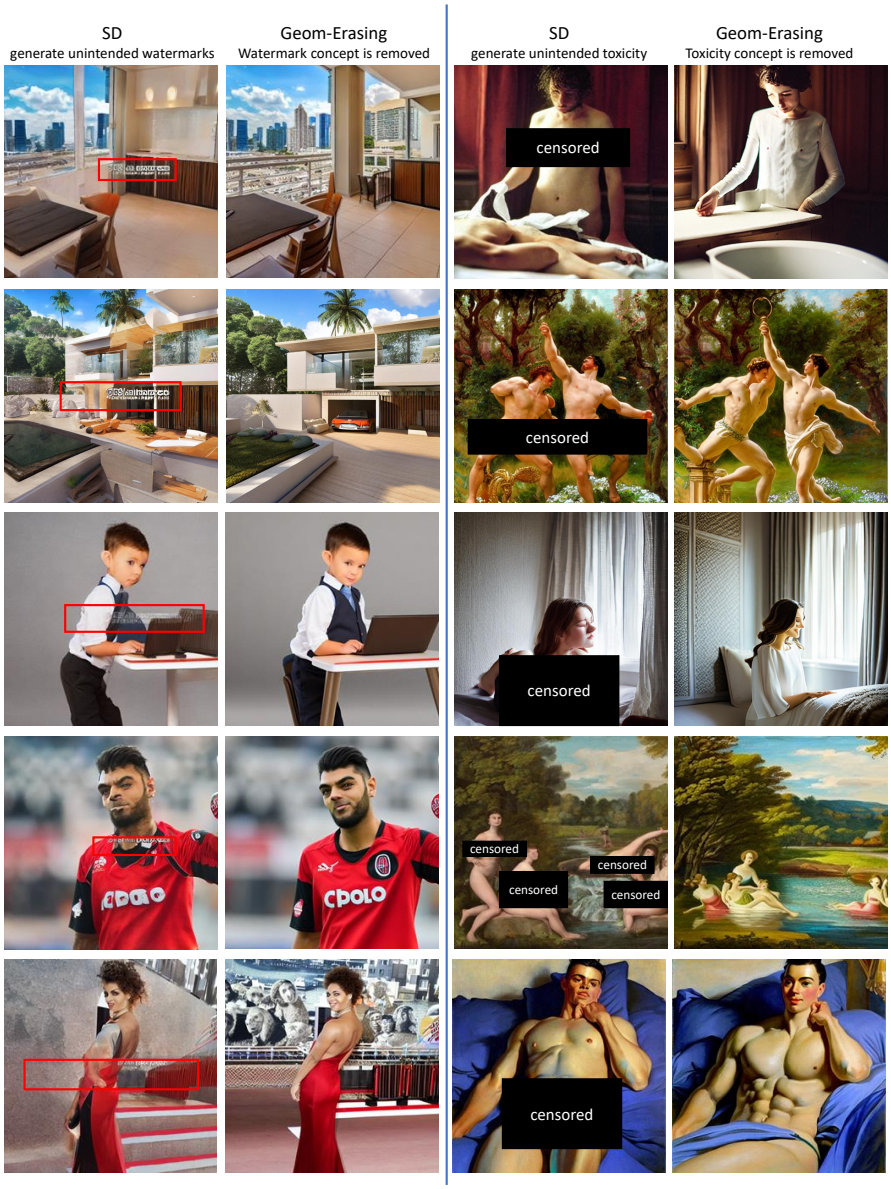


Fig. A3: Erasing Implicit Concepts from SD. We successfully remove watermark and toxicity concepts from generated images while retaining other contents. Optimal viewing is recommended in color and at an enlarged scale.

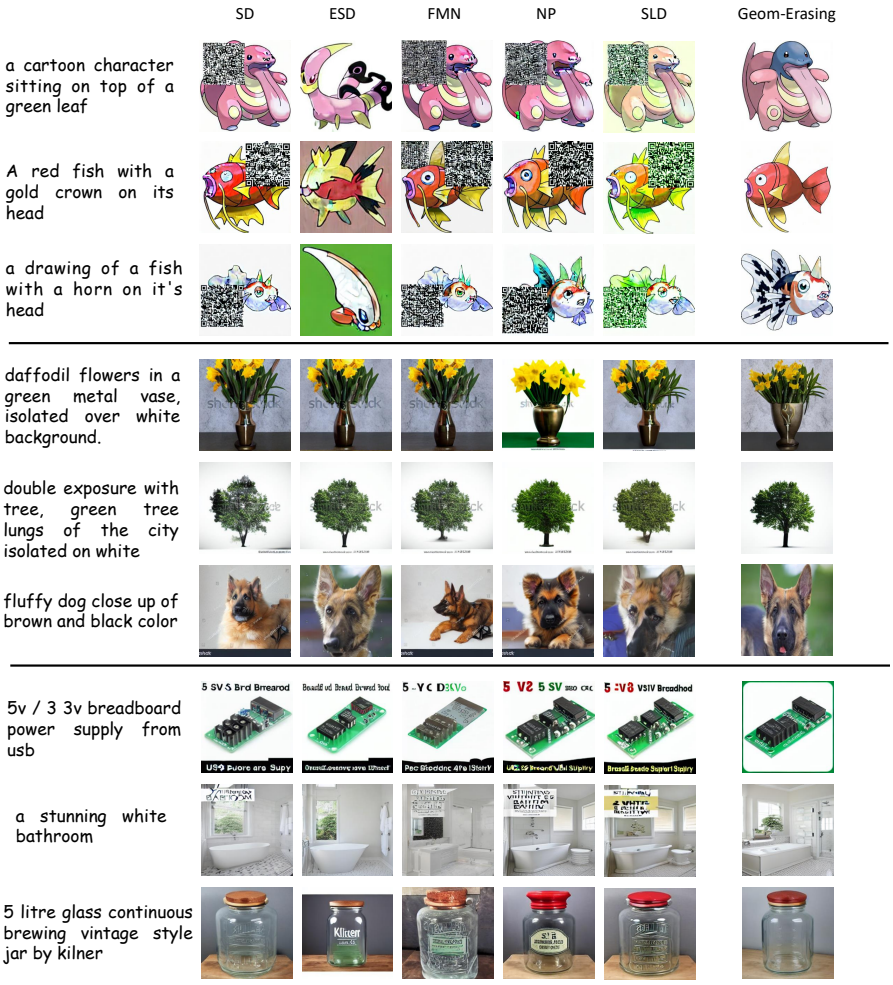


Fig. A4: Erasing implicit concept in ICD. The first group of images are fine-tuned on ICD-QR. The middle and the bottom are fine-tuned on ICD-watermark and ICD-Text, respectively.