Figure 5: **Image samples from our Implicit Concept (IC) dataset.**

APPENDIX

A DATASETS DETAILS

IC-QR. Real QR codes to the Pokemon dataset (Pinkney, 2022). 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[3].

IC-Watermark. Images are collected from CC12M (Changpinyo et al., 2021), 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.

IC-Text. Text images are gathered from LAION (Schuhmann et al., 2021). The training dataset we used is provided by (Yang et al., 2023), 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 Du et al. (2021) is used to detect text within the images.

B LIMITATION OF BASELINE MODELS.

As discussed in Sec.5.2, methods like FMN, NP, and SLD heavily rely on the model's capacity to recognize specific concepts. Following our experiments in ablation, we provide the "watermark" concept as an illustration here to check the cross attention scores to see this recognition ability. It's important to note that NP and SLD operate during the sampling phase, thus depending on the original SD's capabilities. As shown in Fig. 6, SD, NP, SLD, and FMN do not exhibit attention to the regions containing the ground truth watermark. In contrast, GEOM-ERASING successfully directs attention to the meaningful areas that encompass the watermark. This indicates that learning implicit concepts clearly can help erase them.

ESD exhibits superior erasure results compared to other baseline methods; however, it does come at the expense of image quality. This phenomenon is attributed to ESD's tendency to steer generation away from the fine-tuned SD model. While this strategy reduces the occurrence of unwanted implicit concepts, it also affects the generation of other valuable content. As a result, ESD generates fewer images containing implicit concepts but produces images that differ from those generated by other baseline methods, as illustrated in row two, columns 2 and 6 of Fig. 7.

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

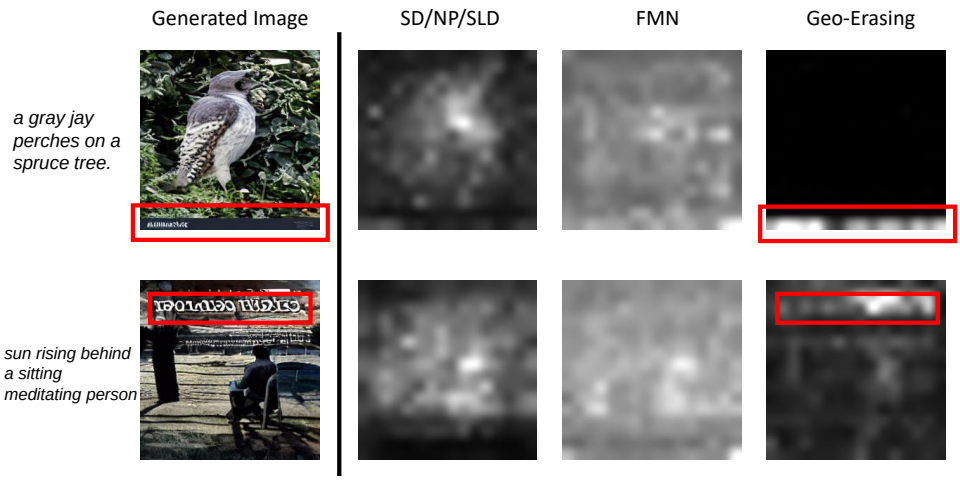

Figure 6: **Visualization of the cross attention map.** The left column is the generated image. Each column in the right part is the cross-attention map between the concept 'watermark' and the image of each method. It can be seen that GEOM-ERASING can attend to the location with watermark, while other methods can not. Since the recognization ability of SD/NP/SLD is poor, they cannot erase the implicit concept.

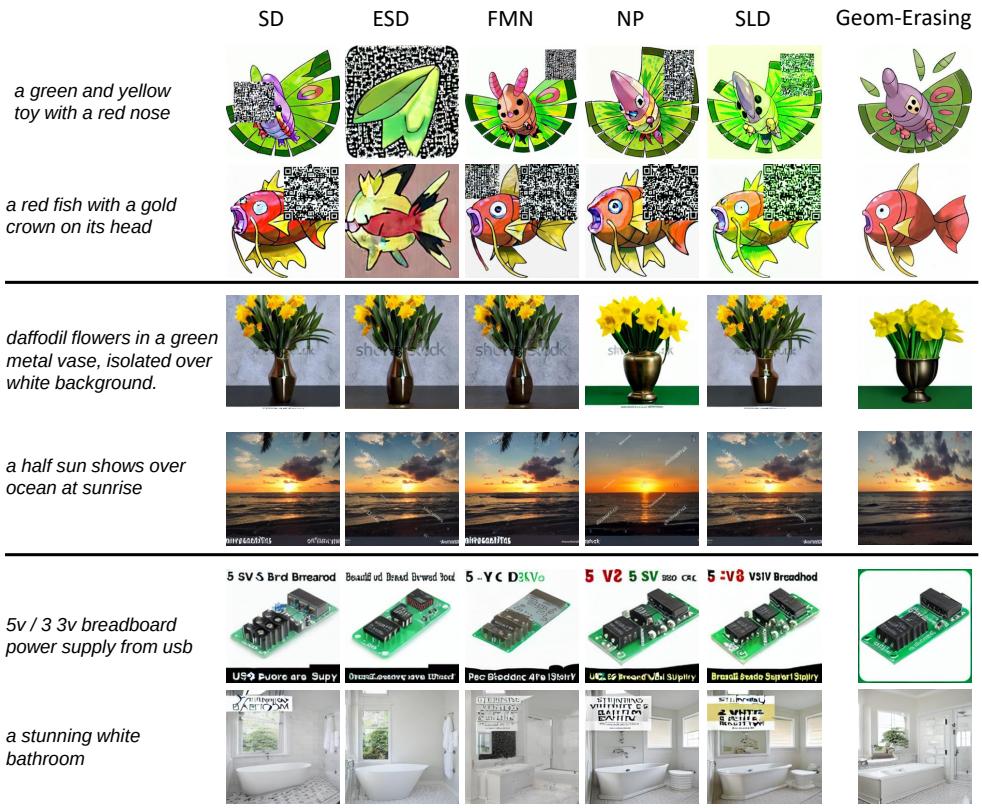

Figure 7: **More generation examples.** The first group of images are fine-tuned on IC-QR. The middle and the bottom are fine-tuned on IC-watermark and IC-Text, respectively.