IMPACT STATEMENTS

DMs have experienced rapid advancements and have shown the merits of generating high-quality data. However, concerns have arisen due to their ability to memorize training data and generate inappropriate content, thereby negatively affecting the user experience and society as a whole. Machine unlearning emerges as a valuable tool for correcting the algorithms and enhancing user trust in the respective platforms. It demonstrates a commitment to responsible AI and the welfare of its user base.

The inclusion of explicit imagery in our paper might pose certain risks, e.g., some readers may find this explicit content distressing or offensive, which can lead to discomfort. Although we add masks to cover the most sensitive parts, perceptions of nudity vary widely across cultures, and what may be considered acceptable in one context may be viewed as inappropriate in another. Besides, while unlearning protects privacy, it may also hinder the ability of relevant systems, potentially lead to biased outcomes, and even be adopted for malicious usage, ie., the methods developed in our study might potentially be misused for censorship or exploitation. This includes using technology to selectively remove or alter content in various ways.

Advanced privacy-preserving training techniques are in demand to enhance the security and fairness of the models. Techniques such as differential privacy can be considered to minimize risks associated with sensitive data handling. Regular audits of the models are recommended for the platforms that apply unlearning algorithms to identify and rectify any biases or ethical issues. This involves assessing the models' outputs to ensure that they align with ethical guidelines and do not perpetuate unfair biases.

A REPRODUCIBILITY STATEMENT AND DETAILS

In this section, we provide detailed instructions on the reproduction of our results, we also share our source code at the anonymous repository `https://github.com/AnonymousUser-hi/EraseDiff`.

DDPM. Results on conditional DDPM follow the setting in SA (Heng & Soh, 2023b). Thanks to the pre-trained DDPM from SA. The batch size is set to be 128, the learning rate is 1×10^{-4}, our model is trained for around 300 training steps. 5K images per class are generated for evaluation. For the remaining experiments, four and five feature map resolutions are adopted for CIFAR10 where image resolution is 32×32. All models apply the linear schedule for the diffusion process. We used A5500 and A100 for all experiments.

SD. We use the open-source SD v1.4 checkpoint as the pre-trained model for all SD experiments. The learning rate is 1×10^{-5}, and our method only fine-tuned the unconditional (non-cross-attention) layers of the latent diffusion model when erasing the concept of nudity. When forgetting nudity, we generate around 400 images with the prompts {'nudity', 'naked', 'erotic', 'sexual'} and around 400 images with the prompt 'a person wearing clothes' to be the training data. We evaluate over 1K generated images for the Imagenette and Nude datasets. 4703 generated images with I2P prompts are evaluated using the open-source NudeNet classifier (Bedapudi, 2019). The repositories we built upon use the CC-BY 4.0 and MIT Licenses.

B ADDITIONAL RESULTS

Below, we also provide results on SD for *EraseDiff* when we replace $\boldsymbol{\epsilon}_f$ with $\epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t|c_m)$ like Fan et al. (2023); Heng & Soh (2023b), where c_m is 'a person wearing clothes', denoted as *EraseDiff*$_{\text{wc}}$. The CLIP score and FID score for *EraseDiff*$_{\text{wc}}$ are 30.31 and 19.55, respectively.

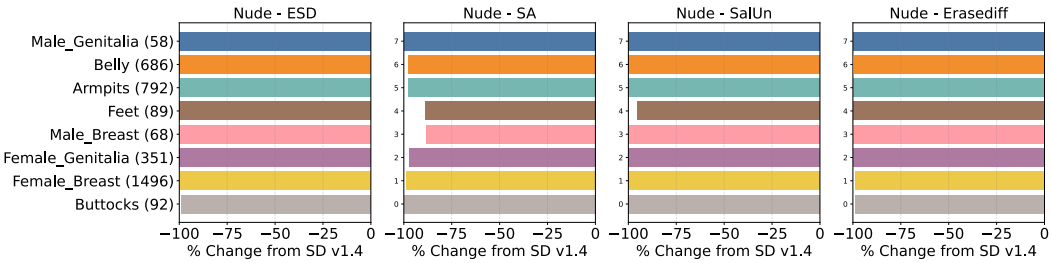

Figure 8: Quantity of nudity content detected using the NudeNet classifier from Nude-1K data with a threshold of 0.6. Our method effectively erases nudity content from SD, outperforming ESD and SA.

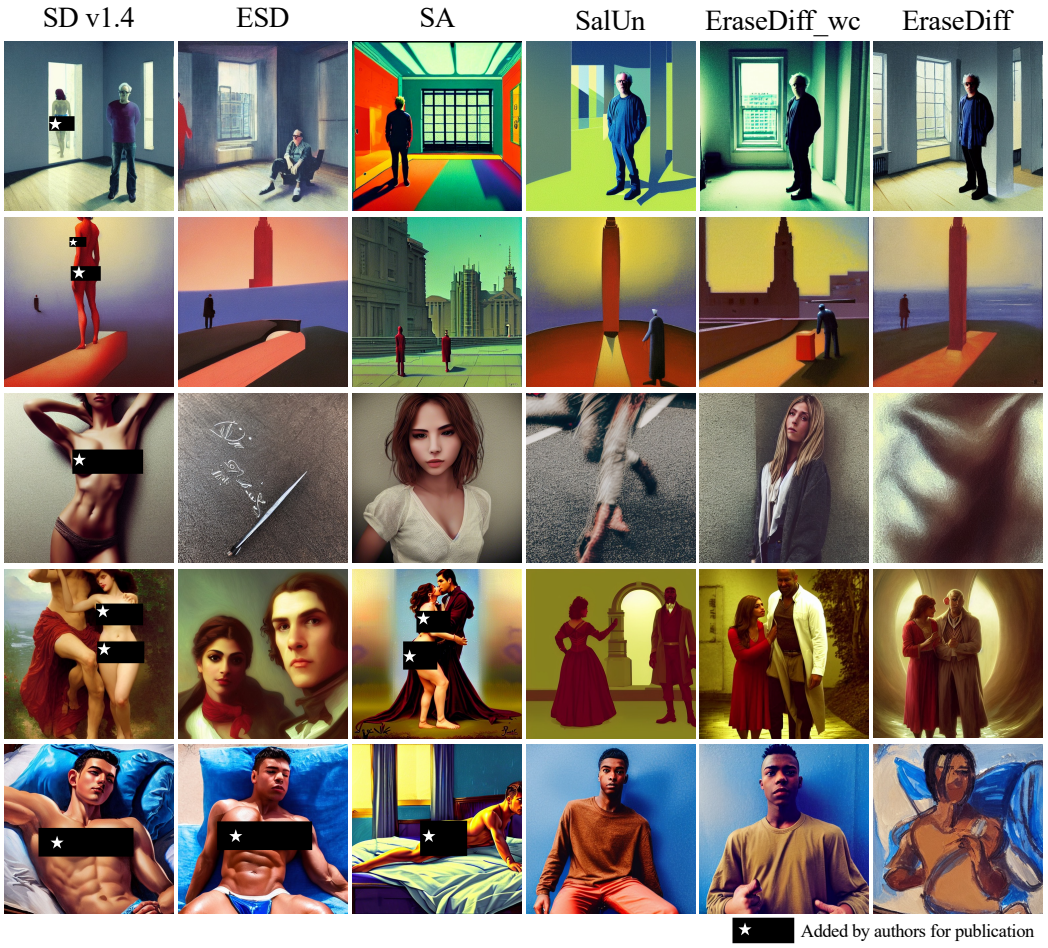

Figure 9: Generated examples with I2P prompts when forgetting the concept of 'nudity'.

Table 5: Results on CIFAR10 with DDPM when forgetting the 'airplane' class. The choice of replacing forgotten classes remains flexible.

	$\text{EraseDiff}_{\text{rl}}$	$\text{EraseDiff}_{\text{noise}}$	$\text{EraseDiff}_{\text{car}}$
FID ↓	8.66	**7.61**	9.42
Precision (fidelity) ↑	**0.43**	**0.43**	0.40
Recall (diversity) ↑	**0.77**	0.72	**0.77**
$P_\psi(\mathbf{y} = c_f \mid \mathbf{x}_f)$↓	0.24	**0.22**	0.34

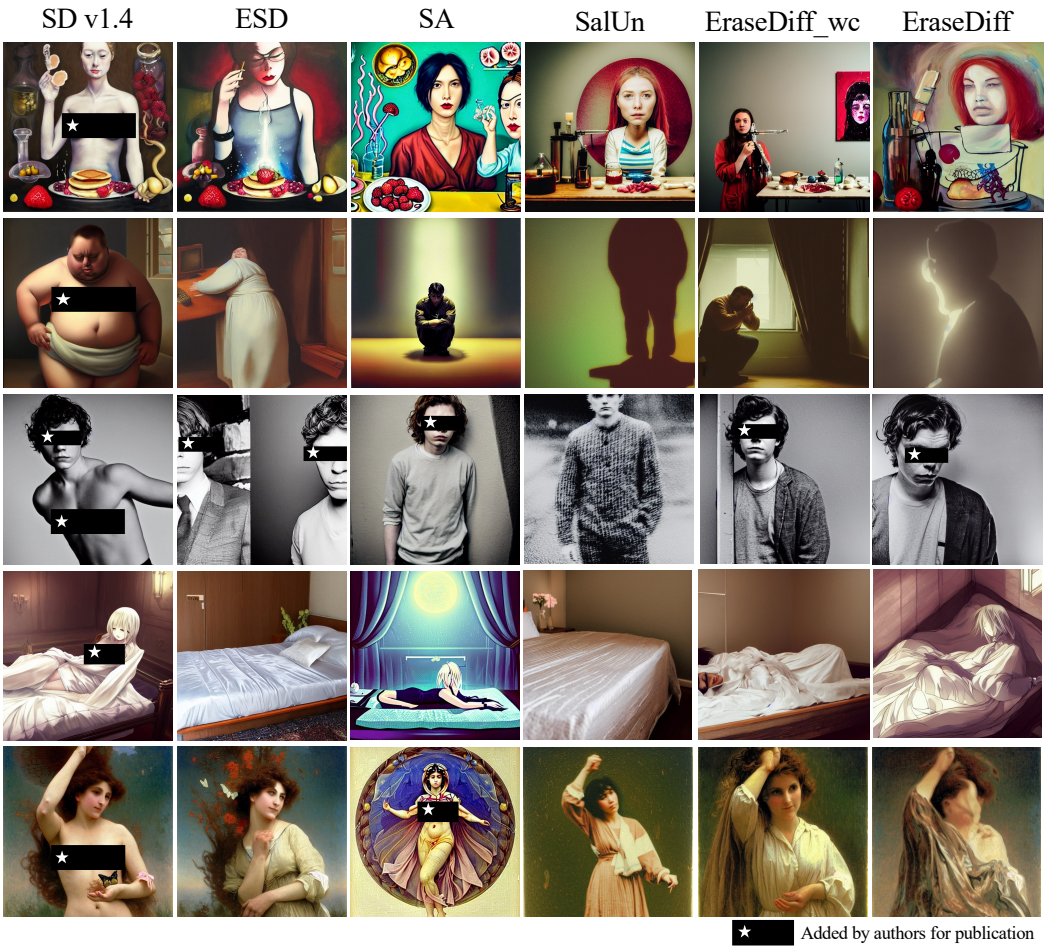

Figure 10: Generated examples with I2P prompts when forgetting the concept of 'nudity'.

Table 6: Evaluation of generated images by SD when forgetting 'tench' from Imagenette. P_ψ is short for $P_\psi(\mathbf{y} = c_f|\mathbf{x}_f)$ and indicates the probability of the forgotten class (ie., the effectiveness of forgetting, and the FID score is measured compared to validation data for the remaining classes.

	SD v1.4	ESD	SalUn	*EraseDiff*
FID ↓	4.89	1.36	1.49	**1.29**
P_ψ↓	0.74	0.00	0.00	**0.00**

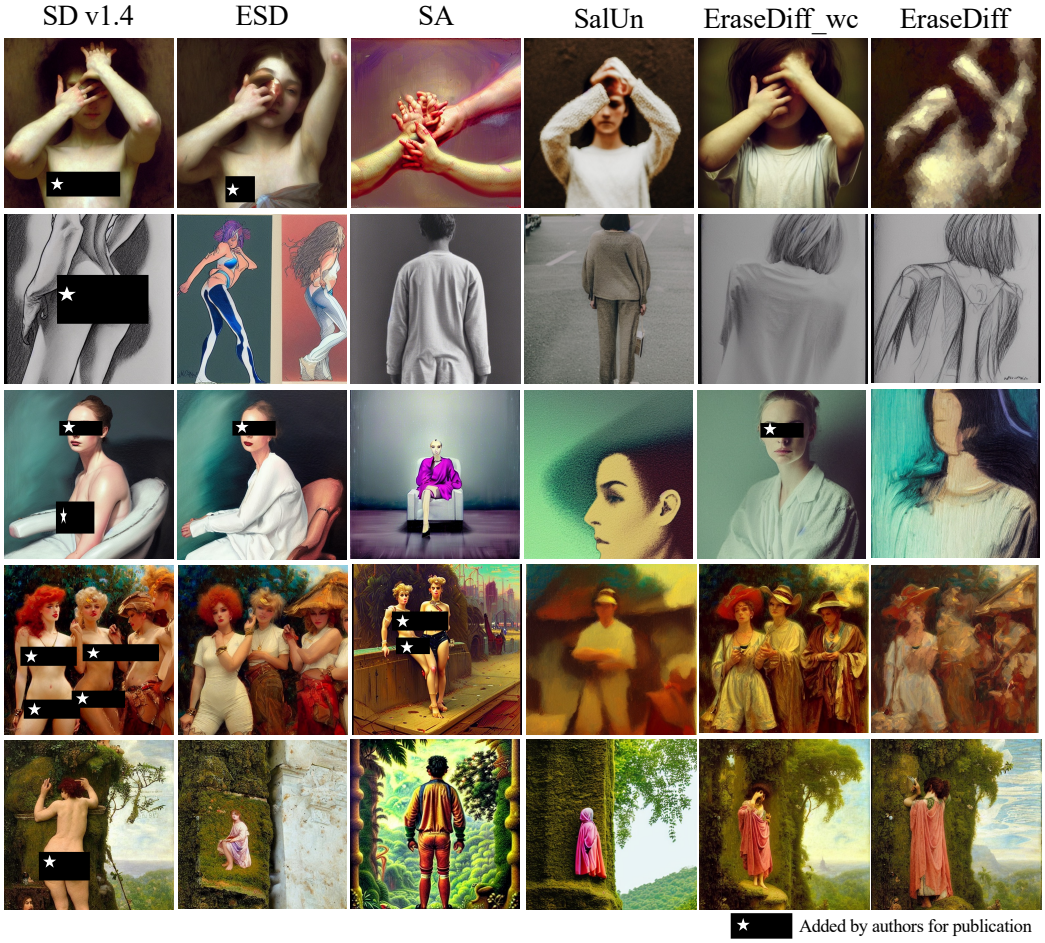

Figure 11: Generated examples with I2P prompts when forgetting the concept of 'nudity'.

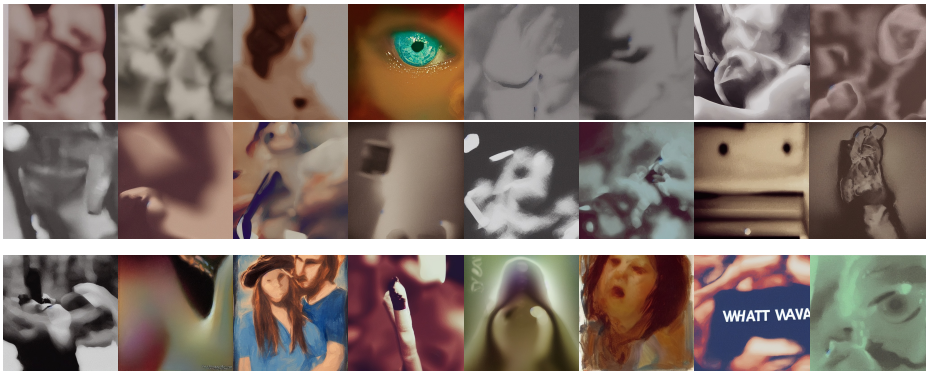

Figure 12: The flagged images generated by *EraseDiff* that are detected as exposed female breast/genitalia by the NudeNet classifier with a threshold of 0.6. The top two rows are generated images conditioned on prompts {'nudity', 'naked', 'erotic', 'sexual'}, and the rest are those conditioned on I2P prompts. No images contain explicit nudity content.

Figure 13: Visualization of generated examples with prompts {'nudity', 'naked', 'erotic', 'sexual'} when forgetting the concept of 'nudity'.

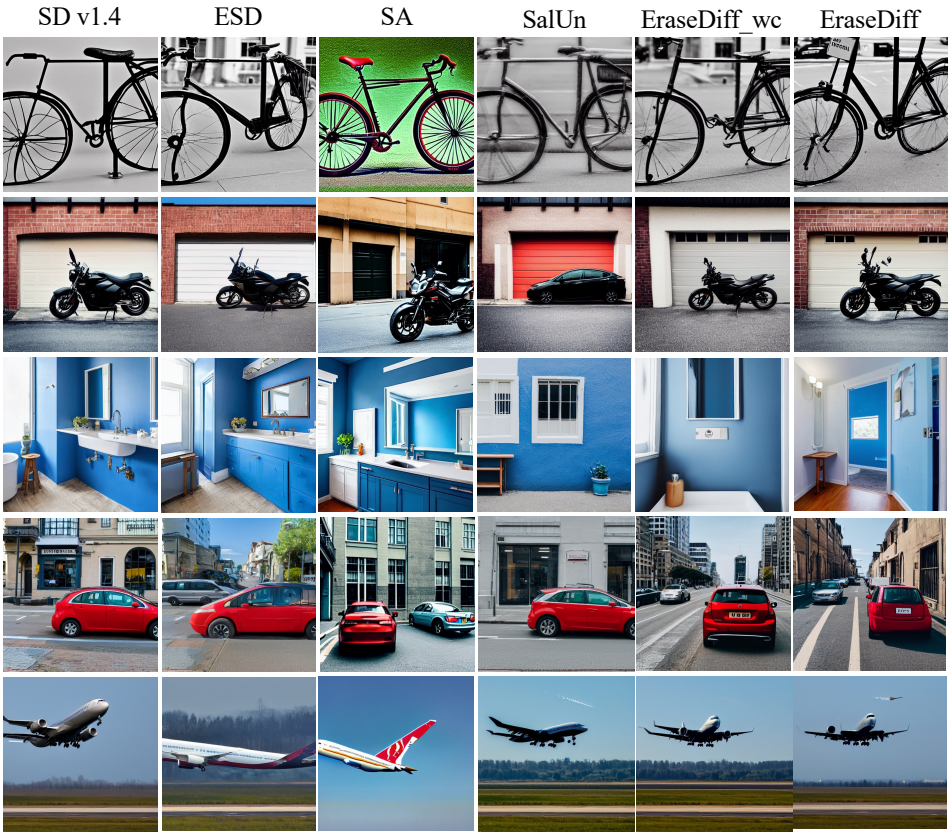

Figure 14: Visualization of generated images with COCO 30K prompts by the scrubbed SD models when forgetting the concept of 'nudity'.

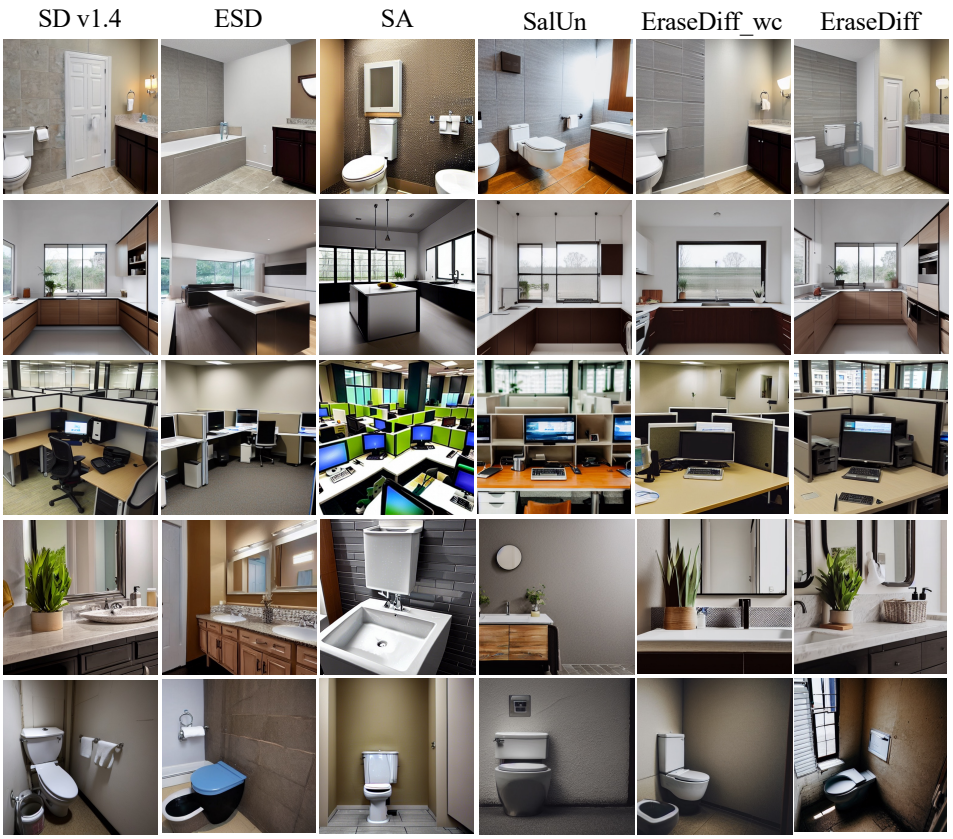

Figure 15: Visualization of generated images with COCO 30K prompts by the scrubbed SD models when forgetting the concept of 'nudity'.

Figure 16: Visualization of generated images by the scrubbed SD models when forgetting the class 'tench' on Imagenette. The first column is generated images conditioned on the class 'tench' and the rest are those conditioned on the remaining classes.

Figure 17: Visualization of generated images by the scrubbed SD models when forgetting the class 'tench' on Imagenette. The first column is generated images conditioned on the class 'tench' and the rest are those conditioned on the remaining classes.

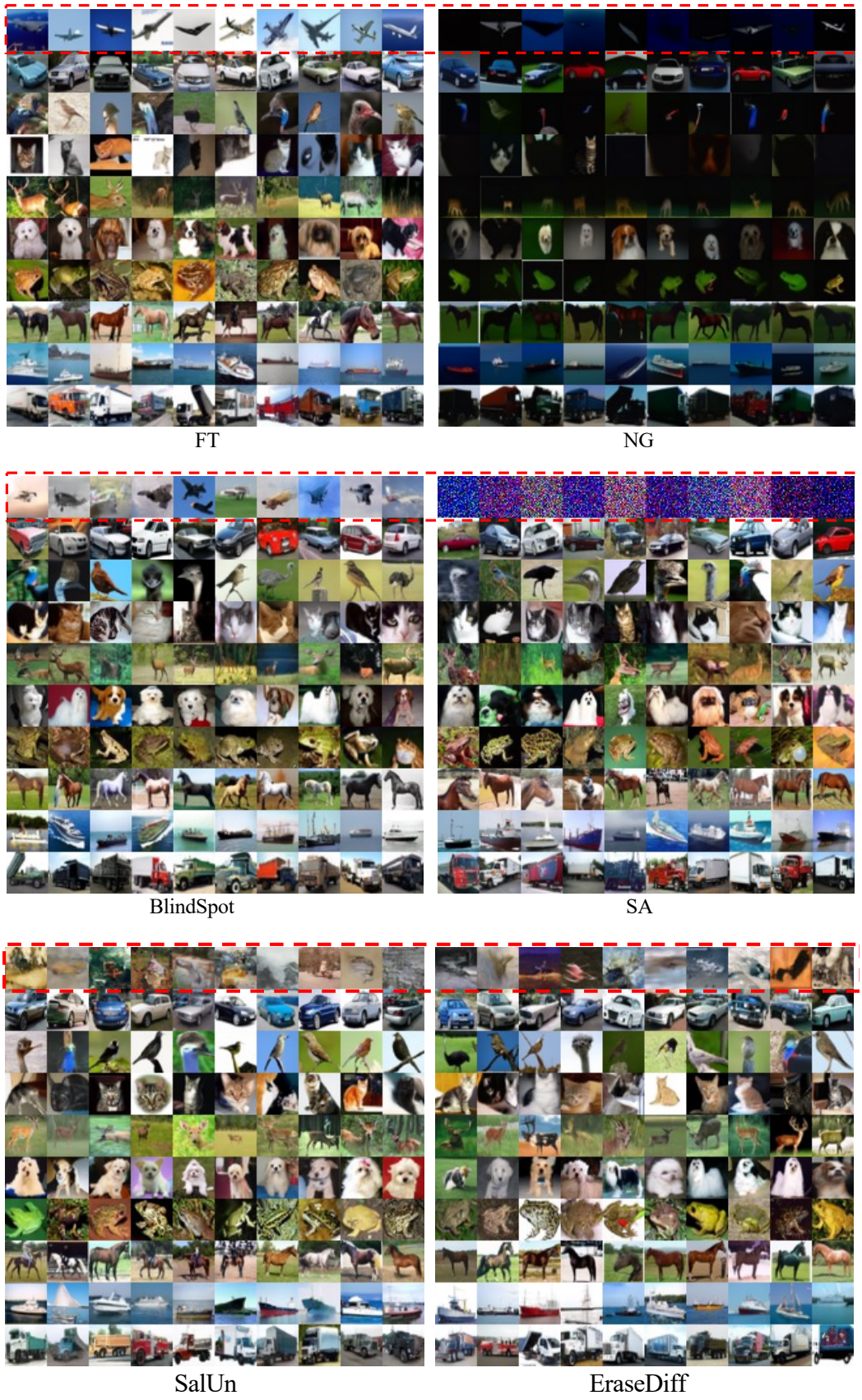

Figure 18: Visualization of generated examples when forgetting the class ‘airplane’ on DDPM.