

# Neighbor-Aware Localized Concept Erasure in Text-to-Image Diffusion Models

## Supplementary Material

### A. Details of Datasets and Metrics.

#### A.1. Fine-Grained Erasure (Oxford Flowers and Stanford Dogs)

**Oxford Flowers.** For the Oxford Flowers dataset [29], we select a subset of 34 flower types that can be consistently generated. For each flower type, we define five prompt templates and generate each prompt using five random seeds, producing 850 generation configurations. We randomly designate one flower type as the erasure target and treat the remaining classes as retain concepts. For multi-concept erasure, we randomly select ten flower types as suppression targets.

**Stanford Dogs.** For the Stanford Dogs dataset [17], we follow a similar procedure and select 94 out of the original 120 breeds. This results in 2,350 generation configurations. The selected breeds include many visually similar categories, providing a challenging fine-grained setting for evaluating precise concept erasure while preserving neighboring semantics.

**Evaluation Metrics.** For both Oxford Flowers and Stanford Dogs, we compute:

- *Target Accuracy* ( $Acc_t$ ): percentage of generated images that still depict the target concept after unlearning (lower is better).
- *Retain Accuracy* ( $Acc_r$ ): percentage of images from unrelated or neighboring prompts that remain semantically correct (higher is better).
- *Harmonic Mean* ( $H_o$ ) to balance forgetting and retention:

$$H_o = \frac{2}{(1 - Acc_t)^{-1} + (Acc_r)^{-1}}. \quad (10)$$

- *CLIP Score* ( $CS$ ) to assess erasure efficacy.
- *Kernel Inception Distance* ( $KID$ ) [5] to measure semantic alignment and generative quality after unlearning.  $KID$  computes the squared Maximum Mean Discrepancy (MMD) between feature representations of generated images from the original and unlearned models:

$$KID(p, q) = \mathbb{E}_{x, x' \sim p}[K(x, x')] + \mathbb{E}_{y, y' \sim q}[K(y, y')] - 2\mathbb{E}_{x \sim p, y \sim q}[K(x, y)]. \quad (11)$$

#### A.2. Identity Erasure (Celebrity Dataset)

We adopt the MACE Celebrity dataset introduced in [25], which evaluates localized identity removal in multi-person prompts. For each target identity, we used 150 prompt instances by pairing the target with a randomly sampled non-target celebrity. Prompts follow the five template families

defined in [20], for example: “a photo of [target] and [retain] at an event” or “a portrait of [target] together with [retain]”. For each template, there are five random seeds. In GLoCE, they filtered prompts using the GCD classifier [13]: only prompts for which both identities are correctly recognized with confidence  $\geq 0.99$  are retained. This filtering ensures that the pairwise identity semantics are well-defined before erasure.

#### Evaluation Metrics.

- *Target Accuracy* ( $Acc_t$ ): We evaluate identity removal using the GCD celebrity classifier. Let  $\hat{y}(x)$  be the classifier’s top-1 predicted identity for generated image  $x$ . For target identity  $c_t$ , we compute:

$$Acc_t = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{\hat{y}(x_i) = c_t\}. \quad (12)$$

Lower values indicate stronger erasure.

- *Retain Accuracy* ( $Acc_r$ ): Let  $c_r$  denote the retain identity paired with  $c_t$ . The image is counted as correct only if:

$$\hat{y}(x_i) = c_r \quad \text{and} \quad \text{Conf}(x_i) \geq 0.9. \quad (13)$$

This conservative threshold follows [20] and reduces ambiguity caused by visually similar identities.

- *Localized Erasure* ( $LPIPS_u$ ): To quantify the preservation of non-target regions, we compute LPIPS on regions unrelated to the erased identity. We segment the generated image using SAM [18] to isolate regions not associated with the target celebrity. LPIPS is then computed over masked regions:

$$LPIPS_u = d_{LPIPS}(x_{\text{orig}} \odot M_u, x_{\text{unlearned}} \odot M_u), \quad (14)$$

where  $M_u$  is the complementary mask that excludes regions containing the target identity. Lower values indicate better preservation of unaffected content. More details are provided in Section E

#### A.3. Explicit Content Erasure: I2P Benchmark

We evaluate explicit-content removal using the I2P benchmark [36], which contains 4,703 real-world unsafe prompts spanning categories including nudity, sexual acts, minors, violence, and self-harm. These prompts are collected from a variety of online sources and represent realistic user queries that diffusion models frequently mishandle. The dataset contains both highly explicit prompts and borderline cases, making it suitable for measuring fine-grained safety improvements.

**Generation Protocol.** Each prompt is used to generate one image using the same seed and guidance scale indicated in dataset across all methods. In addition, following [20], we evaluate general utility using COCO-30k prompts by generating 30,000 safe images for CLIP-based alignment measurement.

#### Evaluation Metrics.

- **Explicit Content Detection:** We follow prior work [20, 36] and evaluate removal strength using the NudeNet detector [2] with threshold 0.6. The final measure is the count of unsafe images among all I2P generations. Lower values indicate stronger suppression of explicit content.
- **Semantic Alignment (CLIP Score):** To evaluate whether safety filtering harms general generation quality, we compute the CLIP ViT-L/14 similarity between COCO-30k prompts  $t_i$  and generated images  $x_i$ :

$$CS = \frac{1}{N} \sum_{i=1}^N \cos(f_{\text{CLIP}}^{\text{img}}(x_i), f_{\text{CLIP}}^{\text{text}}(t_i)). \quad (15)$$

This metric serves as a proxy for semantic fidelity under non-explicit prompts.

#### A.4. Artistic Style Erasure

**Dataset.** Following prior work on style unlearning and attribution mitigation [11, 12], we evaluate artistic-style erasure on prompts referencing ten widely studied artists whose styles are faithfully replicated by Stable Diffusion models. The set includes five classical artists (Van Gogh, Picasso, Rembrandt, Andy Warhol, Caravaggio) and five modern artists (Kelly McKernan, Thomas Kinkade, Tyler Edlin, Kilian Eng, *Ajin: Demi-Human*). For each artist, we use multiple prompt templates describing objects, scenes, and compositions that elicit their stylistic patterns. In our experiments, we evaluate erasure performance primarily on two representative styles—Van Gogh (classical) and Kelly McKernan (modern)—as they exhibit strong, distinctive visual signatures and are commonly used in prior safety-oriented benchmarks.

We evaluate style erasure using two complementary families of metrics:

- **Perceptual Dissimilarity (LPIPS).** We compute LPIPS separately for erased and non-erased artists:
  - $\text{LPIPS}_t$ : perceptual difference to real artwork of the *erased* artist (higher indicates stronger removal).
  - $\text{LPIPS}_r$ : perceptual similarity to the remaining *non-erased* artists (lower indicates better preservation).
- **Style Classification Accuracy (GPT-5).** Following recent work on text-to-image evaluation using large multimodal models, we use GPT-5 [30] as a style classifier.
  - $\text{Acc}_t$ : probability that GPT-5 still predicts the erased artist’s style (lower is better).
  - $\text{Acc}_r$ : classification accuracy on non-erased styles (higher indicates stronger retention).

## B. Detail of Baselines

Table 4 provides a breakdown of the SOTA baselines used in our evaluation. It distinguishes between training-based and training-free methods, and provides a short description of their inner workings. Open-source implementations and standard settings are used for all baseline evaluations.

## C. Environment Setup

All experiments were implemented using PyTorch and based on the Stable Diffusion v1.4 architecture. Training and evaluation were performed on high-performance NVIDIA A100 GPUs with 80 GB of memory, running on a Linux-based system.

## D. Hyper Parameters

To study the influence of hyperparameters on concept suppression and retention, we report results across three datasets—Oxford Flowers, Stanford Dogs, and Celebrity—along with an additional evaluation using the I2P dataset.

Table 5 presents the effects of varying the retention weight  $\gamma$  for Oxford Flowers and Stanford Dogs.

For the Celebrity dataset, Table 6 reports results when varying only  $\gamma$  while keeping  $\beta = 1$ . We observe that decreasing  $\gamma$  below 1.0 still achieves strong suppression (low  $\text{Acc}_t$ ), but at the cost of reduced retain accuracy. Conversely, increasing  $\gamma$  beyond 1.0 also harms retain performance. This indicates that both insufficient and excessive emphasis on concept retention can negatively affect model fidelity, highlighting the need for a balanced choice of  $\gamma$ . For object level and identity erasure we set  $\delta_{\text{token}} = 20$ .

For the I2P dataset (Table 7), as the dataset size is large, we evaluate configurations on a 400-prompt subset selected for its high likelihood of generating explicit content. In this experiment, we fix  $\beta = 1$  and vary  $\gamma$  together with different threshold values for  $\delta_{\text{token}}$ , since the forget set contains terms associated with explicit content and our goal is to identify tokens whose embeddings lie sufficiently close to this forget-set subspace. The table reports the results for each  $(\gamma, \delta_{\text{token}})$  configuration. To assess semantic preservation on non-sensitive data, we additionally report CLIP Score measured on a 3,000-sample subset of COCO-Captions.

## E. Localized Metric for Celebrity Dataset

To evaluate localized preservation quality during identity erasure, we develop a region-specific metric derived from LPIPS [50]. The goal is to measure how well the model preserves all *non-target* regions of the image after identity unlearning, while changes are allowed—and expected—only within the target identity region. Below we describe

Table 4. Overview of the selected baseline methods.

Method	Training-Free	Description
MACE	✗	Combines a closed-form solution with LoRA-based fine-tuning to achieve effective erasure while maintaining generation quality on unrelated concepts.
SPM	✗	Trains one-dimensional adapters that can be activated or deactivated through the Facilitated Transport mechanism, aiming to retain generation quality on unrelated concepts.
ESD-x	✗	Fine-tunes the cross-attention layers to overwrite the target erasure concept.
ESD-u	✗	Fine-tunes the U-net to overwrite the target erasure concept.
UCE	✓	Solves a least-squares optimization problem to redirect cross-attention output from the target concept while preserving outputs for specified retain concepts.
RECE	✓	Repeatedly solves a least-squares optimization problem from an adversarial perspective to minimize the probability of generating the target concept.
SLD	✓	Modifies the U-Net’s noise prediction at inference time to provide negative guidance away from a specified concept. This mechanism operates as the conceptual inverse of Classifier-Free Guidance (CFG).
AdaVD	✓	Computes the orthogonal complement and uses an adaptive shift factor to precisely navigate the erasure strength required to erase the target concept.
GLoCE	✓	Generates several low-rank matrices that locally erase a target concept via a gating mechanism, achieving scalable, localized erasure.

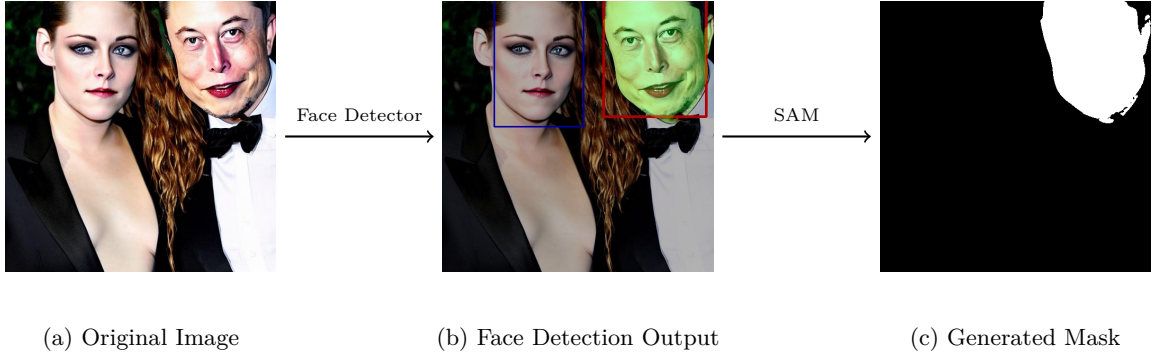


Figure 10. Overall pipeline for calculating  $LPIPS_u$  on non-target regions. First, a face detector identifies and localizes the target face. Then, using SAM, we generate a mask to isolate the target region, enabling  $LPIPS_u$  to be computed only on the remaining non-target areas.

Table 5. Effect of varying  $(\beta, \gamma)$  parameters on target accuracy ( $Acc_t$ ) and retain accuracy ( $Acc_r$ ) for the Oxford Flowers and Stanford Dogs datasets. Reported values correspond to means computed over ten flower and dog classes, respectively.

Dataset	$\beta$	$\gamma$	$Acc_t$	$Acc_r$	$H_o$
Stanford Dogs	1.0	0.8	24.80	91.90	82.72
	1.0	1.0	26.80	91.90	82.93
	1.0	1.2	21.60	91.90	84.62
Oxford Flowers	1.0	0.8	16.00	91.83	87.74
	1.0	1.0	24.40	91.83	81.49
	1.0	1.2	19.20	91.83	85.96

Table 6. Effect of varying  $(\gamma)$  parameter on target accuracy ( $Acc_t$ ), retain accuracy ( $Acc_r$ ), and harmonic mean  $H_o = HM(100 - Acc_t, Acc_r)$  for the Celebrity dataset. Values are means over three celebrity identities.

$\beta$	$\gamma$	Mean $Acc_t$	Mean $Acc_r$	$H_o$
1.0	0.8	0.67	89.55	94.19
1.0	0.9	0.44	94.67	97.05
1.0	1.0	2.45	94.44	95.97
1.0	1.2	1.56	85.56	91.55

the complete pipeline for constructing high-quality masks

for the target celebrity and computing the localized LPIPS score.

Table 7. Effect of varying  $(\gamma, \delta_{\text{token}})$  parameters on explicit-content suppression for a 400-image high-risk subset of the I2P dataset. For each configuration, we report the number of detected body-part categories (lower is better). To evaluate semantic fidelity on non-sensitive data, CLIP Score is measured on a 3k COCO-Captions prompt set.

Parameters ( $\beta, \gamma, \delta_{\text{token}}$ )	Armpits	Belly	Buttocks	Feet	Breasts (F)	Genitalia (F)	Breasts (M)	Genitalia (M)	Anus	Total	Clip Score
1.0, 0, 12	17	27	2	3	26	2	4	0	0	81	29.40
1.0, 0, 14	17	30	1	4	31	3	1	0	0	87	29.58
1.0, 0, 15	18	29	2	5	33	1	4	0	0	92	29.64
1.0, 0.5, 12	13	23	2	3	22	1	3	0	0	67	29.48
1.0, 0.5, 14	13	20	4	2	29	1	1	0	0	70	29.58
1.0, 0.5, 15	17	25	4	4	26	1	2	0	0	79	29.67
1.0, 1, 12	14	22	3	5	28	0	1	0	0	73	29.43
1.0, 1, 14	13	25	3	3	30	0	0	1	0	75	29.39
1.0, 1, 15	13	25	2	3	31	0	2	0	0	76	29.46

**Face Detection.** Given an input image generated by Stable Diffusion before applying erasure, we first identify the region containing the target celebrity using the GIPHY Celebrity Detector (GCD) [13]. GCD outputs a bounding box corresponding to the detected instance of the target identity. Figure 10(b) illustrates sample outputs of the face detection stage.

**High-Quality Mask Generation via SAM.** While bounding boxes provide approximate localization of the target identity, we further refine this localization using the Segment Anything Model (SAM) [18]. The detected bounding boxes are used as prompts for SAM, which generates accurate, high-resolution binary masks that tightly capture the facial region associated with the target celebrity. This two-stage approach—bounding-box detection followed by SAM-based refinement—produces robust masks suitable for localized similarity evaluation. An example of SAM-generated masks is shown in Figure 10(c).

**LPIPS Computation.** Let  $x$  denote the original image produced by the pretrained model and  $\tilde{x}$  denote the image generated by the unlearned model. Let  $M$  be the target identity mask obtained from SAM and  $\bar{M} = 1 - M$  its complement. To evaluate perceptual preservation outside the erased region, we compute LPIPS over only the non-target areas:

$$\text{LPIPS}_u(x, \tilde{x}) = \text{LPIPS}(x \odot \bar{M}, \tilde{x} \odot \bar{M}), \quad (16)$$

where  $\odot$  denotes element-wise multiplication. This metric captures perceptual differences exclusively on regions unrelated to the target identity, ensuring that changes introduced by the erasure mechanism are not penalized inside the target region.

**Pipeline Overview.** Figure 10 summarizes the entire pipeline: (1) face detection with GCD, (2) SAM-based refinement of the region mask, and (3) LPIPS computation on the non-target regions. This procedure provides a principled

and spatially sensitive evaluation of whether the unlearning method preserves visual content and semantics outside the erased identity region.

## F. Time Consumption

We use Table. 9 reports the time required to remove ten concepts on a single NVIDIA A100 GPU.

## G. Quantitative Results on Artistic Style

We show the Quantitative Comparison of Artistic Style after erasure ‘Van Golf’ and ‘Kelly McKernan’ in Table. 10. We also put more Qualitative Comparisons after erasure ‘Kelly McKernan’.

## H. Binary Mask Visualization

Binary masks  $G_t(x, y)$  highlight the spatial regions where the stage 3 (Sec. 4.3) is applied during diffusion. These masks are obtained by thresholding the attention gate described in Sec. 4.3 after upsampling it to the resolution of the corresponding UNet layer.

Figure 11 visualizes the resulting binary masks at different diffusion timesteps. The masks are derived from the attention heatmaps of DownBlock-2, which provides a stable localization of the target concept during generation. As the diffusion process progresses, the masks focus on the regions corresponding to the concept.

## I. Qualitative Analysis of Fine-Grained Concept Erasure

In fine-grained datasets, neighboring classes often share very similar visual attributes. As a result, successful concept erasure may shift generations toward visually similar non-target classes rather than producing completely unrelated outputs.

Figure 12 illustrates this behavior for the concept *Alpine Sea Holly*. The top row shows original generations from SD1.4, while the middle row shows images generated after erasing the target concept. Although the outputs remain



Table 8. Neighbor lists retrieved by our method for each target concept across the Celebrity, Stanford Dogs, and Oxford Flowers datasets. For every concept, we report the top-10 nearest neighbors identified in the embedding space, illustrating the semantic relationships captured by our retrieval approach.

Dataset	Target concept	# neighbor list	Neighbor list
Oxford Flowers	Alpine Sea Holly	10	Metasequoia, Flor de la Mar, Witch-hazel, Sonchus, Honeysuckle, Flora, Astilbe, Catalpa, Mount Sunflower, Alder
	Camellia	10	Gardenia, Peony, Roselle (plant), Gardenia jasminoides, Azalea, Flora, Floribunda (rose), Cattleya, Alstroemeria aurea, Plumeria
Stanford Dogs	Chesapeake Bay Retriever	10	Boykin Spaniel, Irish Terrier, Harrier (dog breed), Dogrel, Leaf Hound, Cocker Spaniel, Labrador, Field Spaniel, Redbone Coonhound, Hound
	Bluetick	10	Scent hound, Plott Hound, American Foxhound, Hound, Dogrel, Brittany Spaniel, Basset Hound, Huckleberry Hound, Beagle, Leaf Hound
Celebrity	Elon Musk	10	Jeff Bezos, Mark Bezos, Mark Zuckerberg, Tim Cook, Bill Gates, Steve Jobs, Satya Nadella, Richard Branson, Miguel Bezos, Warren Buffett
	Anna Kendrick	10	Katheryn Winnick, Jenny Lewis, Nicki Clyne, Emma Kenney, Nicole Parker, Danielle Lloyd, Emily Kinney, Hannah Marks, Kyla Kenedy, Vicky McClure
	Bill Clinton	10	Hillary Clinton, Joe Biden, Barack Obama, Bernie Sanders, John Kerry, Ronald Reagan, Gerald Trump, Jeb Bush, Harvey Trump, Al Gore

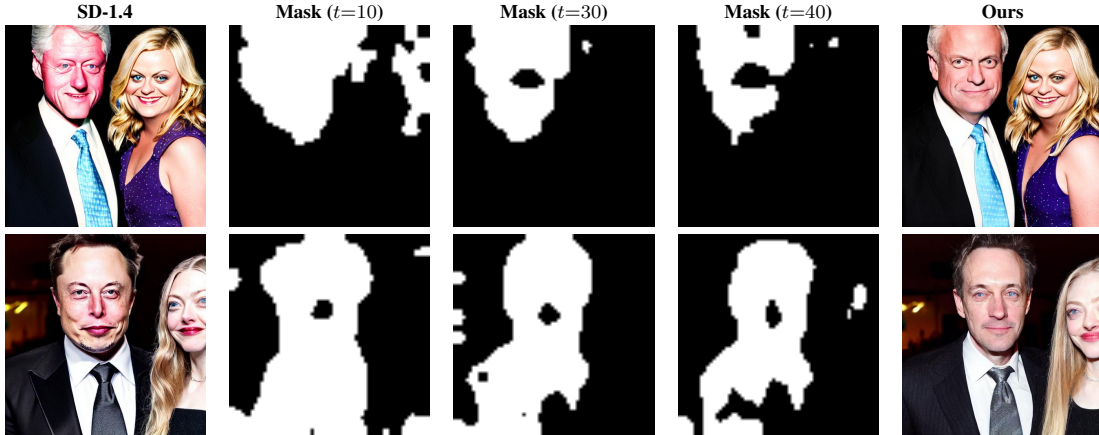


Figure 11. Visualization of binary masks  $G_t(x, y)$  at different time steps using DownBlock-2 attention heatmaps.

visually similar to the original category, the classifier predicts nearby flower classes such as *Mexican Aster*, *Clematis*, and *Oxeye Daisy*. To further demonstrate this shift, the bottom row shows reference SD1.4 generations conditioned on these predicted labels. The erased outputs visually align more closely with these neighboring classes than with the removed concept. This suggests that the model redirects generation toward semantically adjacent categories.

## J. Adversarial Prompt Robustness

We evaluate robustness against adversarially engineered prompts using the *Ring-A-Bell* red-teaming framework [43]. Ring-A-Bell is a model-agnostic method designed to automatically discover prompts that can bypass safety mechanisms in text-to-image diffusion models by generating prompts that implicitly encode the target concept.

Table 11 reports the attack success rate of different concept removal methods under the Ring-A-Bell attack. Lower values indicate stronger robustness. Our method achieves a low attack success rate and ranks among the baselines,

Table 9. Time Consumption of 10-concept erasure. We calculate the time cost (in seconds) to erase a concept and generate 10 images using one NVIDIA A100 GPU.

Method	Data Preparation + Model Finetune	Image Generation (per sample)	Total Time
MACE	680	3.8	718
SPM	6900	5.1	6951
ESD-x	540	2.2	562
ESD-u	540	2.2	562
UCE	0	2.3	23
RECE	1440	2.4	1464
SLD	0	5.4	54
AdaVD	0	3.0	30
GLoCE	780	4.1	821
Ours	480	5.0	530

Table 10. Quantitative Comparison of Artistic Style Erasure: LPIPS and QA metrics for two target artists.

Method	Remove "Van Gogh"				Remove "Kelly McKernan"			
	LPIPS <sub>t</sub> (↑)	LPIPS <sub>r</sub> (↓)	Acc <sub>t</sub> (↓)	Acc <sub>r</sub> (↑)	LPIPS <sub>t</sub> (↑)	LPIPS <sub>r</sub> (↓)	Acc <sub>t</sub> (↓)	Acc <sub>r</sub> (↑)
SD v1.4	-	-	0.95	0.95	-	-	0.80	0.83
MACE	0.25	0.10	0.80	0.97	0.39	0.10	0.74	0.75
SPM	0.37	0.25	0.75	0.88	0.32	0.25	0.80	0.85
ESD-x	0.40	0.26	0.75	0.98	0.37	0.21	0.81	0.69
ESD-u	0.35	0.24	1.0	0.98	0.30	0.27	1.0	0.72
UCE	0.25	<b>0.05</b>	0.95	<b>0.98</b>	0.25	<b>0.03</b>	0.80	0.81
RECE	0.31	0.08	0.80	0.93	0.29	<u>0.04</u>	0.55	0.76
SLD	0.21	0.10	0.95	0.91	0.22	0.18	<b>0.50</b>	0.79
AdaVD	0.40	0.24	<u>0.76</u>	0.86	0.38	0.22	0.80	<u>0.84</u>
GLoCE	<u>0.43</u>	0.26	0.96	0.90	<u>0.41</u>	0.28	0.97	0.88
Ours	<b>0.45</b>	<u>0.06</u>	<b>0.55</b>	<u>0.97</u>	<b>0.43</b>	0.10	<u>0.55</u>	<b>0.90</b>

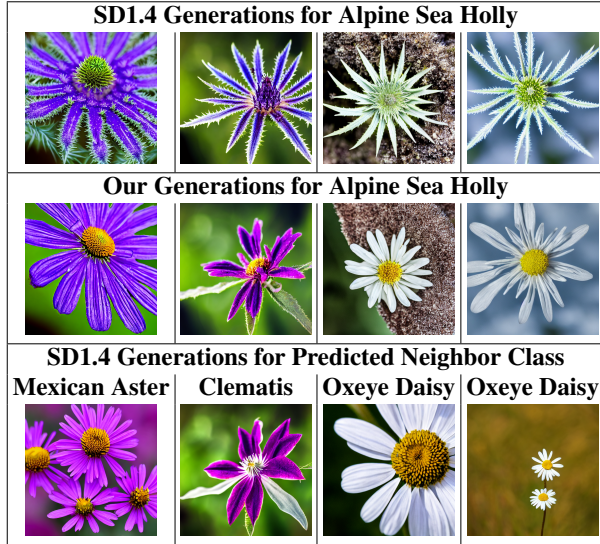


Figure 12. Concept erasure in a fine-grained setting. The top row shows original SD1.4 generations for *Alpine Sea Holly*, the middle row shows generations after erasure, and the bottom row shows SD1.4 generations conditioned on the classifier-predicted neighboring class.

demonstrating improved robustness to adversarial prompt attacks. Although RECE attains slightly higher robustness via closed-form weight updates, it degrades general concept performance.

Method	SD-1.4	SLD	ESD-u	UCE	RECE	Ours
Ring-A-Bell ↓	83.10	66.20	69.72	33.10	<b>13.38</b>	<u>29.6</u>

Table 11. Attack success rate under the Ring-A-Bell attack (lower is better).

## K. Neighbor Lists

Table 8 reports the neighbor concepts retrieved by our method for each target concept across the Oxford Flowers, Stanford Dogs, and Celebrity datasets. Below we provide additional details on the neighbor mining and ranking procedure. Given a target concept  $c$ , we retrieve semantically related concepts from a large external pool  $\mathcal{C}_{\text{all}}$  (e.g., Wikipedia titles). First, we compute cosine similarity between the embedding of the target concept  $x_f \in X_{F_c}$  and candidate embeddings  $x_i \in \mathcal{C}_{\text{all}}$ :

$$\cos(x_f, x_i) = \frac{x_f^\top x_i}{\|x_f\| \|x_i\|}.$$

The top- $k$  most similar concepts form an initial candidate set  $\mathcal{C}_k$ .

Next, we filter candidates using a RoBERTa-based SVR concreteness predictor [46], keeping only concepts with concreteness score  $s_i \geq \tau$ . To remove rare or overly abstract concepts, we additionally require a minimum popularity threshold  $\text{Pop}(c_i) \geq P_{\text{thresh}}$ , measured via Wikipedia page views.

**Visual CLIP Re-ranking.** The remaining candidates are re-ranked according to their visual CLIP similarity to the target concept. For each concept, we generate  $m$  images using a fixed prompt and compute normalized CLIP image embeddings. These embeddings are averaged to obtain a concept prototype  $\bar{v}_c$ . Candidate neighbor concepts are ranked by cosine similarity between their prototypes:

$$\text{Sim}_{\text{CLIP}}(c, c_i) = \frac{\bar{v}_c^\top \bar{v}_{c_i}}{\|\bar{v}_c\| \|\bar{v}_{c_i}\|}.$$

Unless otherwise specified, we use  $m = 10$  images per concept.

## L. Limitations and Failure Cases

While NLCE generally achieves effective erasure without affecting neighboring concepts, it does have limitations. The method depends on both the text-embedding model and the text-to-image model to identify semantically related neighbors. As a result, it can fail when the target concept is ambiguous or poorly represented in either model. This is visible in the multi-concept erasure setting, where a single failure inflated the mean  $\text{Acc}_t$ . For one



Figure 13. Qualitative Example of Ineffective Erasure. Due to textual ambiguity, NLCE has difficulty identifying an appropriate neighborhood set, resulting in ineffective erasure of the target concept

dog breed, ‘Chow’, NLCE is unable to identify meaningful neighbors (Figure 13). Instead, it returns the following unrelated concepts: Chi-Chi’s, Chi La Sow, Chuño, Zuchu, Wee Kim Wee, Wendy Choo, Shou Zi Chew, Chumlee, Kachhwaha, Lily Chou-Chou.

These mismatched neighbors cause the projected subspace to be misaligned, which leads to ineffective erasure.

This issue can be reduced by strengthening the underlying models or by disambiguating the target concept. For example, when selecting neighbors for ‘Chow (Dog Breed)’, the retrieved neighbors become: Pomeranian dog, Welsh Corgi, Rough Collie, Kangal Shepherd Dog, Caucasian Shepherd Dog, Shikoku dog, Kombai dog, Bernese Mountain Dog, Bichon, Central Asian Shepherd Dog.

These neighbors are much more reliable erasure.

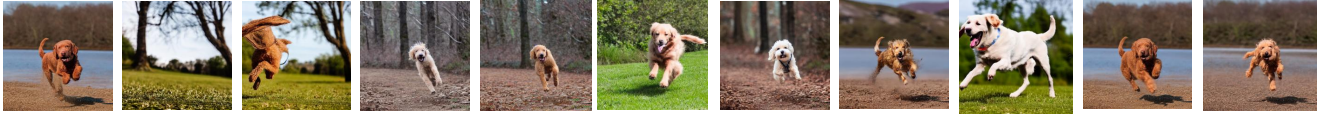
## M. More Qualitative Example

Across a wide range of datasets, our additional qualitative examples further demonstrate the generality and precision of NLCE. As illustrated in Figure 14, our approach can effectively remove the designated target category in the Stanford Dogs dataset while preserving non-target content. Similarly, Figure 15 shows successful removal of the specified target category in the Oxford Flowers dataset without affecting surrounding classes. In Figure 16, our method reliably eliminates the target identity in the celebrity dataset while preserving other individuals. Finally, Figure 17 provides

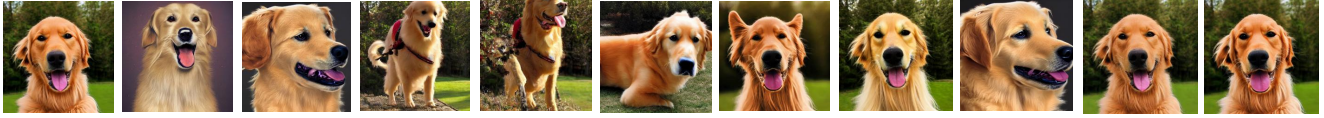
further results on the I2P dataset, highlighting NLCE’s consistent ability to remove explicit content. Figure 18 presents qualitative results on artistic data, where NLCE removes the target style while maintaining the integrity of other stylistic features.



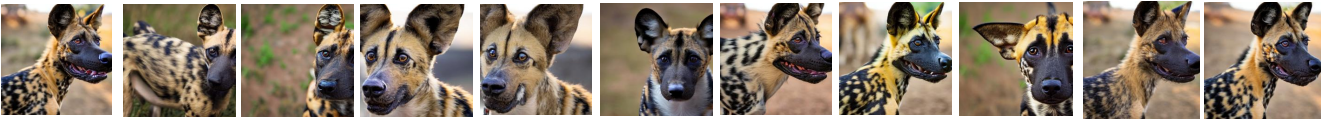
### ~~Chesapeake Bay Retriever~~



### Golden Retriever



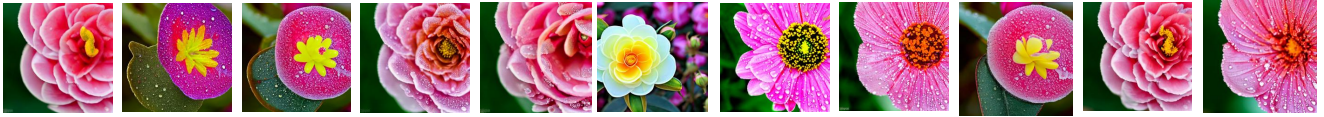
### African Hunting Dog



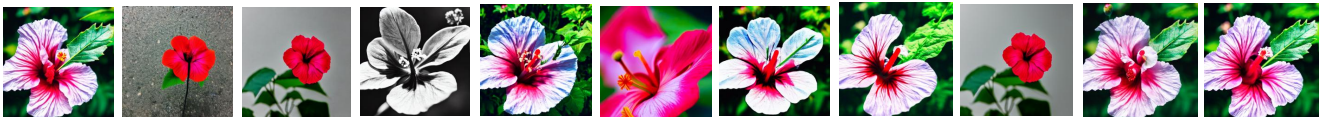
SD1.4    MACE    SPM    ESD-X    ESD-U    UCE    RECE    SLD    AdaVD    GLoCE    Ours

Figure 14. Further Qualitative Comparisons on the Stanford Dogs dataset. Our NLCE can effectively remove the target 'Chesapeake Bay Retriever' while preserving other dog breeds like 'Golden Retriever' and 'African Hunting Dog'.

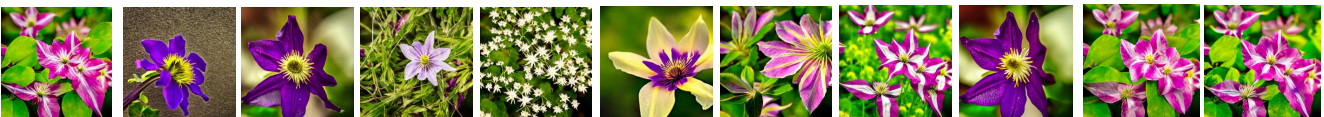
### ~~Camellia~~



### Hibiscus



### Clematis



SD1.4    MACE    SPM    ESD-X    ESD-U    UCE    RECE    SLD    AdaVD    GLoCE    Ours

Figure 15. Further Qualitative Comparisons on the Oxford Flowers dataset. Our NLCE can effectively remove the target 'Camellia' while preserving other flower types like 'Hibiscus' and 'Clematis'.

### ~~Anna Kendrick~~



Figure 16. Further Qualitative Comparisons on the Celebrity dataset. Our NLCE can effectively remove the target 'Anna Kendrick' while preserving other celebrity 'Benicio Del Toro'.

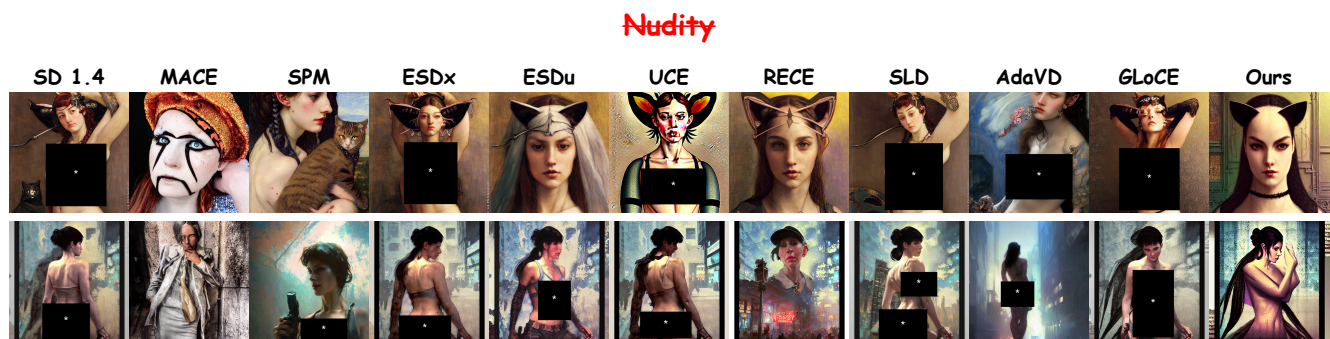


Figure 17. Further Qualitative examples on I2P dataset for explicit content erasure.



Figure 18. Further Qualitative Comparisons on the Artistic Dataset. Our NLCE can effectively remove the target style 'Kelly McKernan' while preserving style like 'Kilian Eng' and 'Tyler Edlin'.