Fine-Grained Erasure in Text-to-Image Diffusion-based Foundation Models

Supplementary Material

1. Theoretical Basis for Concept Lattice

Based on the literature, as noted in the work from Duda et al. [4], we extend the observations of Boiman et al. [2] as a theoretical justification for the proposed nearest neighbor-based concept lattice, which approximates the gold-standard Naive Bayes classifier for constructing the adjacency set.

Theorem 1 (k-NN Approximation to Naive Bayes in $\mathbb{R}^d$)**.** *Let $\mathbf{x} \in \mathbb{R}^{h\times w\times c}$ represent an image with dimensions height h, width w, and channels c. Let the mapping function $\phi : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^d$ project the image $\mathbf{x}$ into a latent feature space $\mathbb{R}^d$, where $d \ll hwc$. Assume that the latent features $z := \phi(\mathbf{x})$ are conditionally independent given the class label $C \in \mathcal{C}$.*

Then, the k-Nearest Neighbors (k-NN) classifier operating in $\mathbb{R}^d$ converges to the Naive Bayes classifier as the sample size $N \to \infty$, the number of neighbors $k \to \infty$, and $k/N \to 0$. Specifically,

$$\lim_{N\to\infty} P\big(C_{k\text{-}NN}(\phi(\mathbf{x})) = C_{NB}(\mathbf{x})\big) = 1. \tag{1}$$

Proof Outline: Let $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ be a dataset consisting of images $\mathbf{x}_i \in \mathbb{R}^{h\times w\times c}$ and their corresponding class labels $y_i \in \mathcal{C}$. Each image $\mathbf{x}_i$ is mapped to a latent space $\mathbb{R}^d$ through the mapping function $\phi : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^d$, resulting in a latent feature vector $z_i := \phi(\mathbf{x}_i)$.
We assume the following:

- The latent feature vectors $\phi(\mathbf{x})$ are conditionally independent given the class label y.
- The representation function $\phi(\mathbf{x})$ preserves the class-conditional structure in $\mathbb{R}^d$, such that images of the same class remain clustered in proximity to one another.
- d is sufficiently large ensuring high separability between classes while remaining lower-dimensional than the original input space, i.e., in $d \ll hwc$.

Proof: We follow the outline above proceeding step by step.
Step 1: Bayes Optimal Classifier (Naive Bayes) The Bayes optimal classifier is defined as the classifier that minimizes the expected classification error by choosing the class that maximizes the posterior probability $P(C = c|\mathbf{x})$. Under the Naive Bayes assumption, the posterior decomposes as follows:

$$P(C = c|\mathbf{x}) = \frac{P(\mathbf{x}|C = c)P(C = c)}{P(\mathbf{x})}. \tag{2}$$

Given the conditional independence of $\mathbf{z}$ in $\mathbb{R}^d$, the class-conditional likelihood $P(\mathbf{x}|C = c)$ is factorized over the components of the latent vector $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_d(\mathbf{x}))$, i.e.,

$$P(\mathbf{x}|C = c) = \prod_{j=1}^{d} P(\phi_j(\mathbf{x})|C = c). \tag{3}$$

Thus, the decision rule of the Naive Bayes classifier becomes:

$$C_{\text{NB}}(\mathbf{x}) = \arg\max_{c\in\mathcal{C}} P(C = c) \prod_{j=1}^{d} P(\phi_j(\mathbf{x})|C = c). \tag{4}$$

Step 2: k-Nearest Neighbor Classifier in $\mathbb{R}^d$
The k-NN classifier operates in the latent space $\mathbb{R}^d$, assigns a class label $C_{\text{k-NN}}$ to a query vector $\phi(\mathbf{x})$ by selecting the nearest instance in $\mathbb{R}^d$. For two images $\mathbf{x}$ and $\mathbf{x}_i$, we can defined it formally as:

$$C_{\text{k-NN}} = \arg\max_{i} \text{sim}(\phi(\mathbf{x}), \phi(\mathbf{x}_i)) = \frac{\langle \phi(\mathbf{x}), \phi(\mathbf{x}_i)\rangle}{\|\phi(\mathbf{x})\|\|\phi(\mathbf{x}_i)\|}. \tag{5}$$

The k-NN classifier assigns the label to the query image $\mathbf{x}$ by aggregating the labels of its k -nearest neighbors $\mathcal{N}_k(\mathbf{x})$ in the latent space. This is formally described as:

$$C_{\text{k-NN}}(\phi(\mathbf{x})) = \arg\max_{c\in\mathcal{C}} \sum_{\mathbf{x}_i\in\mathcal{N}_k(\mathbf{x})} \mathbb{I}(y_i = c), \tag{6}$$

where $\mathbb{I}(y_i = c)$ is the indicator function, returning 1 if $y_i = c$ and 0 otherwise.

Step 3: Convergence of k-NN to Bayes Optimal Classifier
As established by Covert & Hart et al. [3] in statistical learning theory, the k-NN classifier converges to the Bayes optimal classifier as $N \to \infty$, provided that $k \to \infty$ and $k/N \to 0$. That is, for sufficiently large N and k, the decision rule of the k-NN classifier approximates that of the Bayes optimal classifier $C_{\text{Bayes}(\mathbf{x})}$, i.e.,

$$\lim_{N\to\infty} P(C_{\text{K-NN}}(\phi(\mathbf{x})) = C_{\text{Bayes}}(\mathbf{x})) = 1. \tag{7}$$

This convergence holds because, with increase in N, $\mathcal{N}_k(\mathbf{x})$ increasingly reflects the local distribution of data around $\mathbf{x}$, which aligns with the underlying class-conditional probability distribution.

Step 4: Consistency of k-NN with Naive Bayes in $\mathbb{R}^d$
Given the Naive Bayes assumption that the components $\phi_j(\mathbf{x})$ of the latent representation $\phi(\mathbf{x})$ are conditionally independent given the class label, the Bayes optimal classifier in this latent space is precisely the Naive Bayes classifier $C_{\text{NB}}(\mathbf{x})$. Therefore, we have:

$$C_{\text{Bayes}}(\mathbf{x}) = C_{\text{NB}}(\mathbf{x}), \tag{8}$$

where $C_{\text{Bayes}}(.)$ operates on the latent representations $\phi(\mathbf{x})$. Combining equation 7 with the equation 6, we conclude that:

$$\lim_{N\to\infty} P(C_{\text{K-NN}}(\phi(\mathbf{x})) = C_{\text{NB}}(\mathbf{x})) = 1. \tag{9}$$

This establishes that the CLIP-based K-NN classifier converges to the Naive Bayes classifier as the sample size grows, provided the assumptions of conditional independence hold in the latent space $\mathbb{R}^d$.

Remarks:

- In high-dimensional spaces, Bayers et al. [1] proposed concentration of distances implying that Euclidean distance and Cosine similarity perform similarly as $d \to \infty$, ensuring that the use of cosine similarity in latent space provides robust distance-based classification.
- In our implementation, the mapping function $\phi : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^d$ is a pre-trained CLIP model, serving for dimensionality reduction where $d \ll hwc$.
- The CLIP model's latent space captures abstract and semantic features, reducing the dependency between the components of $\phi(\mathbf{x})$. This makes the assumption of conditional independence more plausible in $\mathbb{R}^d$, allowing Naive Bayes to model the class-conditional likelihoods accurately in the latent space.
- For k-NN to converge to optimal Bayes classifier, k must satisfy $k \to \infty$ and $N \to \infty$.

---

2. Adjacency Inflection Analysis

This section examines the breaking point of existing algorithms in preserving adjacency—specifically, at what similarity threshold these methods begin to fail. To evaluate robustness, we analyze the performance of each algorithm as semantic similarity increases, using fine-grained classes from ImageNet-1k and other fine-grained datasets. Figure 1 illustrates the relationship between CLIP-based semantic similarity (circular axis, %) and average adjacency accuracy (radial axes).

Results show that FMN and ESD degrade significantly at 78% similarity, while Receler fails at 80%. Although SPM demonstrates moderate resilience, it begins to falter beyond 90% similarity, marking a critical threshold where all existing methods fail to preserve adjacency effectively.

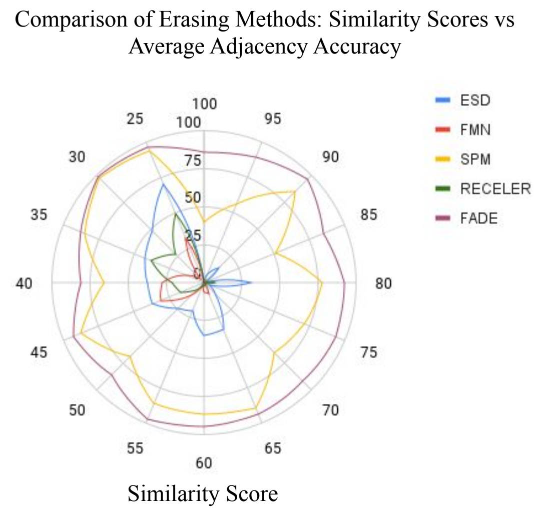

Figure 1. **Radar plot comparing FADE with existing unlearning methods (ESD, FMN, SPM, Receler).** For a fair analysis, methods with $A_{\text{er}} \leq 20\%$ are considered. The plot shows structural similarity scores (circular axis, %) and adjacency accuracy (radial axes) on concepts from the ImageNet-1k dataset. Most methods begin to degrade beyond a similarity score of 70%, with SPM remaining resilient until 90% and FADE demonstrating the highest robustness.

In stark contrast, FADE maintains high adjacency accuracy even at elevated similarity levels, demonstrating superior robustness. These findings validate FADE's efficacy in adjacency-aware unlearning, outperforming state-of-the-art approaches under challenging fine-grained conditions.

3. Adjaceny Retention Analysis

During training, FADE explicitly considers the top-k adjacent classes (with $k = 5$ in all experiments). However, to ensure FADE's generalization beyond the explicitly trained adjacent classes, we evaluate its performance on unseen adjacent concepts (i.e., classes with rank > 5).

We assess FADE's adjacency retention by analyzing classification accuracy across the top-10 adjacent classes for each target concept (as detailed in Table 5). Using classifiers trained on their respective datasets, we measure retention accuracy for Stanford Dogs, Oxford Flowers, and CUB datasets. Figure 2 illustrated a clear trend: as the semantic similarity decreases (from the closest adjacent class A1 to the furthest A10), retention accuracy consistently improves.

To further validate this trend, we extend our analysis to the top-100 adjacent classes per target concept, where the first 5 classes are seen during training, and the remaining 95 are unseen. As shown in Figure 3, FADE consistently maintains retention accuracy above 75% across both seen and unseen adjacent classes, demonstrating its strong generalization capability even after erasure of the target concept.

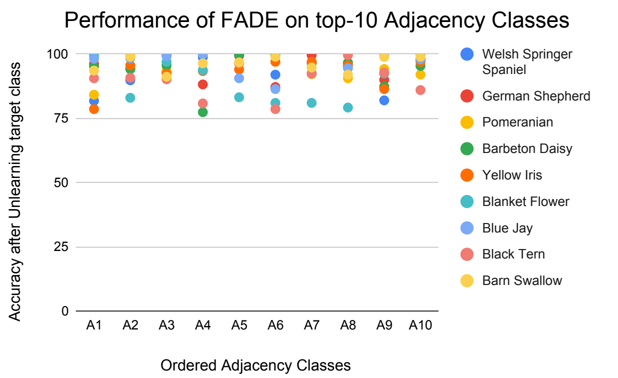

Figure 2. For each target class in Table 5, we illustrate the performance of FADE on the top-10 adjacent classes. Adjacent classes are ordered by similarity scores. It is observed that FADE generalizes well on all adjacent classes after unlearning the target class.

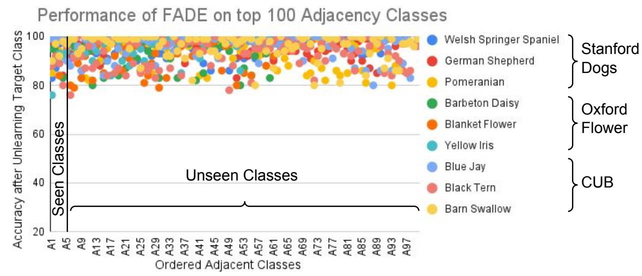

Figure 3. Extended experiment for Adjaceny Retention from 5 unseen concepts to 95 unseen concepts. We observe that the performance remains consistent for all unseen classes.

4. Extended Quantitative Results

As previously discussed, we utilize Stanford Dogs, Oxford Flower, and CUB datasets to evaluate the proposed FADE and existing state-of-the-art algorithms. We present the adjacency set with their similarity scores in Table 5.

We report the classification accuracy for each class in the adjacency set of each target class from the Stanford Dogs, Oxford Flowers, and CUB datasets in Tables 2, 3, and 4. These results extend the findings reported in Table 1 of the main paper. The original model (SD) has not undergone any unlearning, so higher accuracy is better. The remaining models are comparison algorithms, and for each of them, the model should achieve lower accuracy on the target class to demonstrate better unlearning and higher accuracy on neighboring classes to show better retention of adjacent classes. From Tables 2, 3, and 4, it is evident that FADE effectively erases the target concept while preserving adjacent ones, outperforming the comparison algorithms by a significant margin across all three datasets, followed by SPM and CA. This demonstrates the superior capability at erasure and retention of the proposed FADE algorithm.

5. Extended Qualitative Results

To provide a focused and detailed view of the results, Figures 4 and 5 (borrowed from the main paper) visualize the performance of various unlearning algorithms.

Figure 4, 6, 7, and 8 present the generation results for one target class and its adjacency set from each dataset, before and after applying unlearning algorithms. The first row in each of these figures displays images generated by the original Stable Diffusion (SD) model, followed by outputs from each unlearning method. Consistent with the quantitative results in Table 1 (main paper), ESD, FMN, and Receler fail to retain fine-grained details of neighboring classes. CA and SPM perform slightly better, retaining general structural features, but they struggle with specific attributes such as color and texture, especially in examples like dog breeds (e.g., Brittany Spaniel, Cocker Spaniel), bird species (e.g., Florida Jay, Cardinal), and flower species. These methods often result in incomplete erasure of the target concept or poor retention of neighboring classes.

In contrast, FADE achieves a superior balance by effectively erasing the target concept while preserving the fine-grained details of related classes, as demonstrated by the sharper distinctions in the adjacency sets. FADE's capability is further evaluated on ImageNet-1k for target classes such as Balls, Trucks, Dogs, and Fish. Table 5 lists the neghiboring classes identified using Concept Lattice to construct the adjacency set for each target class. Notably, adjacency sets generated by Concept Lattice closely align with the manually curated fine-grained class structures reported by Peychev et al. [8], validating the accuracy and reliability of Concept Lattice.

As shown in Table 2 of the main paper, FADE outperforms all baseline methods, achieving at least a 12% higher ERB score compared to SPM, the next-best algorithm. FMN and CA exhibit poor performance in both adjacency retention and erasure tasks, highlighting the robustness of FADE in fine-grained unlearning scenarios.

Further, human evaluation results for FADE and baseline algorithms are presented in Table 1, capturing erasing accuracy (A_{er}) and average adjacency retention accuracy ($\hat{A}_{\text{adj}}$). We also capture their balance through the proposed Erasing-Retention Balance (ERB) score.

According to human evaluators, Receler achieves the highest A_{er} (86.66%) but fails in adjacency retention, with $\hat{A}$adj close to zero, resulting in a minimal ERB score (0.06). FMN and CA show suboptimal performance, with FMN favoring erasure and CA favoring retention, yielding ERB scores of 43.07 and 38.43, respectively.

FADE outperforms all baselines with the highest ERB score (59.49), balancing effective erasure (A_{er} of 51.94%) and strong adjacency retention ($\hat{A}_{\text{adj}}$ of 69.62%). These re-

	A_{er}	$\hat{A}_{\text{adj}}$	ERB
ESD	73.33	37.22	49.38
FMN	49.16	38.33	43.07
CA	30.13	53.05	38.43
SPM	40.83	61.66	49.13
Receler	**86.66**	0.03	0.06
FADE (ours)	51.94	**69.62**	**59.49**

Table 1. **Comparison of FADE with state-of-the-art unlearning methods based on evaluations by human participants.** If prediction of human evaluator is correct, a score of 1 was given; otherwise, a score of 0 was given. The performance is reported as a percentage. According to the user study, FADE effectively balances the erasure of the target concept with the retention of neighboring concepts.

sults highlight FADE's ability to achieve adjacency-aware unlearning without significant collateral forgetting, setting a benchmark for fine-grained erasure tasks.

6. Implementation Details

For all experiments and comparisons, we use Stable Diffusion v1.4 (SD v1.4) as the base model. The datasets constructed (discussed in Section 3.3 of the main paper) are generated using SD v1.4, and the same model is used to generate images for building the Concept Lattice. In Equation 9 (of the main paper), we set the base parameters as λ_{er}: 3.0, λ_{adj}: 1000, λ_{guid}: 50. These values may vary depending on the specific target class being unlearned. For equation 6, value of δ is 1.0 across all experiments. Throughout all experiments, we optimize the model using AdamW, training for 500 iterations with a batch size of 4. All the experiments are performed on one 80 GB Nvidia A100 GPU card.

For all baseline algorithms, we utilize their official GitHub repositories and fine-tune only the cross-attention layers wherever applicable(ESD[6], CA[7]). In the case of CA[7], each target class is assigned its superclass as an anchor concept. For instance, for the Welsh Springer Spaniel, the anchor concept is its superclass, dog. Similarly, for concepts in the Stanford Dogs dataset, the anchor concept is set to dog, while for the Oxford Flowers dataset, it is flower, and for CUB, it is bird. This selection strategy is consistently applied when defining preservation concepts while evaluating UCE.

For calculation of A_{er} and $\hat{A}_{\text{adj}}$ in Table 1 of main paper, we utilize ResNet50 as the classification model. Specifically, we fine-tune ResNet50 on 1000 images generated for each class in Stanford Dogs, Oxford Flowers and CUB datasets. For ImageNet classes in Table 2 and Table 3 of main paper we utilize pre-trained ResNet50. For I2P related evaluations, we utilize NudeNet.

7. Additional Analysis

Choosing an adjacent concept for CA: We conduct an additional experiment using English Springer as the anchor concept for Welsh Springer Spaniel in Concept Ablation [7]. This yields an ERB score of 69.40, significantly lower than FADE's 95.97. While WSS$\rightarrow$ES improves erasure compared to WSS$\rightarrow$Dog, it severely degrades retention ($\hat{A}_{\text{adj}}$=61.4), indicating disruption in the learned manifold.
Adversarial Robustness: To assess FADE's resilience against adversarial prompts, we conducted an experiment using the Ring-a-Bell! [5] adversarial prompt generation algorithm. For Table 1 of the main paper, we evaluated prompts on both the original and unlearned models across Stanford Dogs, Oxford Flowers, and CUB datasets. The target class accuracies (lower is better) for the original model were 92.8, 65.4, and 45.8, while FADE significantly reduced them to 20.8, 1.3, and 5.4, demonstrating strong robustness against adversarial prompts.
Concept Unlearning Induces Concept Redirection: Our experiments reveal an intriguing phenomenon where unlearning a target concept often results in its redirection to an unrelated concept. As illustrated in Figure 9, this effect is particularly evident with algorithms like ESD and Receler. For example, after unlearning the "Blanket Flower," the model generates a "girl with a black eye" when prompted for "Black-eyed Susan flower" and produces an image of "a man named William" for the prompt "Sweet William flower." Similarly, for bird classes such as "Cliff Swallow" and "Tree Swallow," the unlearning process redirects the concepts to unrelated outputs, such as trees or cliffs.

Interestingly, this redirection is primarily observed in algorithms like ESD and Receler, which struggle to maintain semantic coherence post-unlearning. In contrast, SPM and the proposed FADE algorithm demonstrate robust performance, effectively erasing the target concept without inducing unintended redirections, thereby preserving the model's semantic integrity.

References

[1] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *Database Theory—ICDT'99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7*, pages 217–235. Springer, 1999.

[2] Oren Boiman, Eli Shechtman, and Michal Irani. In defense of nearest-neighbor based image classification. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[3] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1): 21–27, 1967.

[4] Richard O Duda, Peter E Hart, and David G Stork. Pattern classification wiley new york, 2001.

		SD (Original)	ESD	FMN	CA	SPM	Receler	FADE (ours)
Target Concept - 1	Welsh Springer Spaniel	99.10	0.00	1.24	37.00	42.27	0.00	0.45
Adjacent Concepts	Brittany Spaniel	95.42	0.00	0.00	58.00	54.60	0.00	81.85
	English Springer	89.72	0.00	0.00	51.00	24.40	0.00	89.86
	English Setter	94.00	0.00	0.00	56.84	73.85	0.00	93.00
	Cocker Spaniel	98.85	40.00	0.00	82.25	84.10	0.00	98.82
	Sussex Spaniel	99.68	60.62	0.00	85.00	88.75	12.00	99.65
Target Concept - 2	German Shepherd	99.62	0.00	0.00	20.85	0.89	0.00	0.00
Adjacent Concepts	Malinois	99.00	0.00	0.00	51.86	57.43	0.00	96.29
	Rottweiler	98.10	0.00	0.25	54.58	70.24	0.00	94.25
	Norwegian elkhound	99.76	10.00	0.00	63.00	59.00	0.00	99.65
	Labrador retriever	95.00	30.00	1.86	73.60	73.65	0.00	88.20
	Golden retriever	99.86	60.00	0.88	75.44	93.81	14.00	99.44
Target Concept - 3	Pomeranian	99.84	0.00	1.85	32.00	66.49	0.00	0.24
Adjacent Concepts	Pekinese	98.27	0.00	0.00	64.63	86.29	0.00	84.24
	Yorkshire Terrier	99.90	40.00	0.00	89.00	98.62	0.64	99.62
	Shih Tzu	98.85	60.00	0.00	92.00	95.65	2.55	93.65
	Chow	100.00	10.00	1.20	87.60	98.22	3.27	100.00
	Maltese dog	99.45	60.00	1.86	88.88	97.45	0.00	96.45

Table 2. **Comparison of Classification Accuracy on Stanford Dogs Dataset**. We compare the classification accuracy (in %) of various models on classes from the Stanford Dogs dataset before and after unlearning on each class in the adjacency set. The original model (SD) has not undergone any unlearning (higher accuracy is better), while the rest are comparison unlearning algorithms. For each algorithm, the model should exhibit lower accuracy on the target class and higher accuracy on the adjacent concepts. It is evident that FADE significantly outperforms all the comparison algorithms.

		SD (Original)	ESD	FMN	CA	SPM	Receler	Ours
Target Concept	Barbeton Daisy	91.50	0.00	24.45	29.89	30.00	0.00	0.12
Adjacent Concepts	Oxeye-Daisy	99.15	20.00	4.20	75.60	86.86	1.85	95.24
	Black Eyed Susan	97.77	70.00	2.20	79.08	94.00	6.00	94.25
	Osteospermum	99.50	50.00	0.85	90.20	94.80	0.60	95.65
	Gazania	93.50	0.00	1.00	62.28	83.84	0.00	77.45
	Purple Coneflower	99.80	100.00	1.65	85.80	98.85	23.22	99.87
Target Concept	Yellow Iris	99.30	0.00	0.00	32.45	51.69	0.00	0.00
Adjacent Concepts	Bearded Iris	85.25	0.00	0.00	20.27	63.60	0.65	78.68
	Canna Lily	98.72	0.00	0.00	54.45	76.48	1.63	95.68
	Daffodil	94.65	10.00	5.25	59.89	88.00	0.00	92.45
	Peruvian Lily	98.50	20.00	0.00	64.00	88.20	0.00	93.45
	Buttercup	98.00	0.00	32.00	78.46	92.23	0.48	94.00
Target Concept	Blanket Flower	99.50	0.00	37.00	73.00	46.00	0.00	0.00
Adjacent Concepts	English Marigold	99.56	0.00	3.00	94.25	98.00	0.00	99.43
	Gazania	93.55	0.00	0.00	66.87	74.24	0.00	83.00
	Black Eyed Susan	97.77	0.00	0.47	72.84	93.45	1.27	97.00
	Sweet William	97.75	0.00	0.68	66.62	70.00	2.45	93.88
	Osteospermum	99.50	20.00	0.25	92.68	86.45	0.83	83.20

Table 3. **Comparison of Classification Accuracy on Oxford Flower Dataset**. We compare the classification accuracy (in %) of various models on classes from the Oxford Flower dataset before and after unlearning on each class in the adjacency set. The original model (SD) has not undergone any unlearning (higher accuracy is better), while the rest are comparison unlearning algorithms. For each algorithm, the model should exhibit lower accuracy on the target class and higher accuracy on the adjacent concepts. It is evident that FADE significantly outperforms all the comparison algorithms.

[5] Tsai et al. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *ICLR*, 2024.

[6] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023.

[7] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023.

[8] Momchil Peychev, Mark Müller, Marc Fischer, and Martin

		SD (Original)	ESD	FMN	CA	SPM	Receler	Ours
Target Concept	Blue Jay	99.85	0.00	0.00	31.42	14.68	0.00	0.00
Adjacent Concepts	Florida Jay	98.55	0.00	1.40	46.25	69.88	0.00	98.24
	White Breasted Nuthatch	99.00	5.00	0.00	72.00	85.08	0.00	98.44
	Green Jay	99.90	10.00	0.26	52.00	85.00	0.00	99.20
	Cardinal	100.00	30.00	0.11	86.85	96.43	3.48	100.00
	Blue Winged Warbler	92.97	20.00	0.54	49.28	64.24	0.00	90.65
Target Concept	Black Tern	86.65	0.00	4.00	22.45	13.87	0.00	0.00
Adjacent Concepts	Forsters Tern	92.35	0.00	2.86	35.29	41.20	0.00	90.65
	Long Tailed Jaeger	97.66	30.00	9.85	81.08	90.67	0.66	90.84
	Artic Tern	89.50	0.00	0.45	22.20	37.85	0.00	90.26
	Pomarine Jaeger	88.29	0.00	0.82	52.80	63.64	0.00	80.85
	Common Tern	98.10	10.00	0.60	78.20	77.46	0.00	96.45
Target Concept	Barn Swallow	99.40	0.00	1.25	57.06	7.48	0.00	0.45
Adjacent Concepts	Bank Swallow	9.79	0.00	0.65	54.60	30.21	0.00	93.60
	Lazuli Bunting	99.75	70.00	0.65	82.00	88.40	0.00	99.00
	Cliff Swallow	93.20	0.00	0.00	77.00	47.63	0.86	91.25
	Indigo Bunting	96.80	70.00	17.46	87.68	93.00	5.22	96.45
	Cerulean Warbler	96.90	50.00	2.65	88.40	89.00	0.45	96.80

Table 4. **Comparison of Classification Accuracy on CUB Dataset**. We compare the classification accuracy (in %) of various models on classes from the CUB dataset before and after unlearning on each class in the adjacency set. The original model (SD) has not undergone any unlearning (higher accuracy is better), while the rest are comparison unlearning algorithms. For each algorithm, the model should exhibit lower accuracy on the target class and higher accuracy on the adjacent concepts. It is evident that FADE significantly outperforms all the comparison algorithms.

Figure 4. **Qualitative comparison between existing and proposed algorithms for erasing target concepts and testing retention on neighboring fine-grained concepts.** Results are shown for one target concept each from the Stanford Dogs, Oxford Flowers, and CUB datasets.

Vechev. Automated classification of model errors on imagenet. *Advances in Neural Information Processing Systems*, 36:36826–36885, 2023.

Stanford Dogs	Welsh Springer Spaniel		German Shepherd		Pomeranian	
Dataset	Class Names	Similarity Score	Class Names	Similarity Score	Class Names	Similarity Score
Adjacent Class - 1	Brittany Spaniel	98.19	Malinois	95.84	Pekinese	92.67
Adjacent Class - 2	English Springer	97.05	Rottweiler	92.52	Yorkshire Terrier	92.42
Adjacent Class - 3	English Setter	95.12	Norwegian Walkhound	92.51	Shih Tzu	91.58
Adjacent Class - 4	Cocker Spaniel	95.10	Labrador Retriever	91.63	Chow	90.99
Adjacent Class - 5	Sussex Spaniel	93.62	Golden Retriever	91.43	Maltese Dog	90.96
Adjacent Class - 6	Blenheim Spaniel	93.05	Collie	90.79	Chihuaha	90.49
Adjacent Class - 7	Irish Setter	92.90	Doberman	90.37	Papillon	90.38
Adjacent Class - 8	Saluki	92.83	Black and Tan Coonhound	90.27	Samoyed	89.22
Adjacent Class - 9	English Foxhound	92.81	Bernese Mountain dog	90.06	Australian Terrier	89.15
Adjacent Class - 10	Gordon Setter	92.35	Border collie	89.56	Toy poodle	89.05

Oxford Flower	Barbeton Daisy		Yellow Iris		Blanket Flower	
Dataset	Class Names	Similarity Score	Class Names	Similarity Score	Class Names	Similarity Score
Adjacent Class - 1	Oxeye Daisy	98.65	Bearded Iris	96.38	English Marigold	95.13
Adjacent Class - 2	Black eyed susan	95.56	Canna Lily	92.48	Gazania	92.38
Adjacent Class - 3	Osteospermum	94.99	Daffodil	92.33	Black Eyed Susan	90.64
Adjacent Class - 4	Gazania	93.91	Peruvian Lily	92.18	Sweet William	90.00
Adjacent Class - 5	Purple Coneflower	93.27	Buttercup	91.55	Osteospermum	89.94
Adjacent Class - 6	Pink yellow dahlia	91.74	Hippeastrum	91.33	Barbeton Daisy	89.86
Adjacent Class - 7	Sunflower	91.18	Moon Orchid	91.08	Purple Coneflower	89.45
Adjacent Class - 8	Buttercup	91.17	Ruby Lipped Cattleya	90.75	Snapdragon	89.42
Adjacent Class - 9	Japanese Anemone	90.56	Hard-Leaved Pocket Orchid	90.36	Wild Pansy	88.06
Adjacent Class - 10	Magnolia	90.27	Azalea	90.02	Pink Yellow Dahlia	88.01

CUB Dataset	Blue Jay		Black Tern		Barn Swallow	
	Class Names	Similarity Score	Class Names	Similarity Score	Class Names	Similarity Score
Adjacent Class - 1	Florida Jay	93.62	Forsters Tern	96.19	Bank Swallow	95.91
Adjacent Class - 2	White Breasted Nuthatch	92.50	Long Tailed Jaeger	95.54	Lazuli Bunting	93.91
Adjacent Class - 3	Green Jay	91.91	Artic Tern	94.79	Cliff Swallow	92.47
Adjacent Class - 4	Cardinal	90.86	Pomarine Jaeger	94.52	Indigo Bunting	92.04
Adjacent Class - 5	Blue Winged Warbler	90.00	Common Tern	93.35	Cerulean Warbler	91.65
Adjacent Class - 6	Downy Woodpecker	88.84	Elegant Tern	93.03	Blue Grosbeak	91.58
Adjacent Class - 7	Indigo Bunting	88.83	Frigatebird	91.32	Tree Swallow	91.33
Adjacent Class - 8	Cerulean Warbler	88.80	Least tern	91.28	Black Throated Blue Warbler	90.43
Adjacent Class - 9	Black Throated Blue Warbler	88.64	Red legged Kittiwake	91.22	Blue winged warbler	89.25
Adjacent Class - 10	Clark Nutcracker	88.39	Lysan Albatross	90.17	White breasted kingfisher	89.16

Table 5. Description of the adjacency set for target classes from the Stanford Dogs, Oxford Flowers, and CUB datasets, along with their similarity scores.

Figure 5. Comparison of FADE with various algorithms for erasing the 'garbage truck' class in Fine-Grained and Coarse-Grained Unlearning. The target class, adjacency set and the retain set and constructed from the ImageNet-1k dataset.

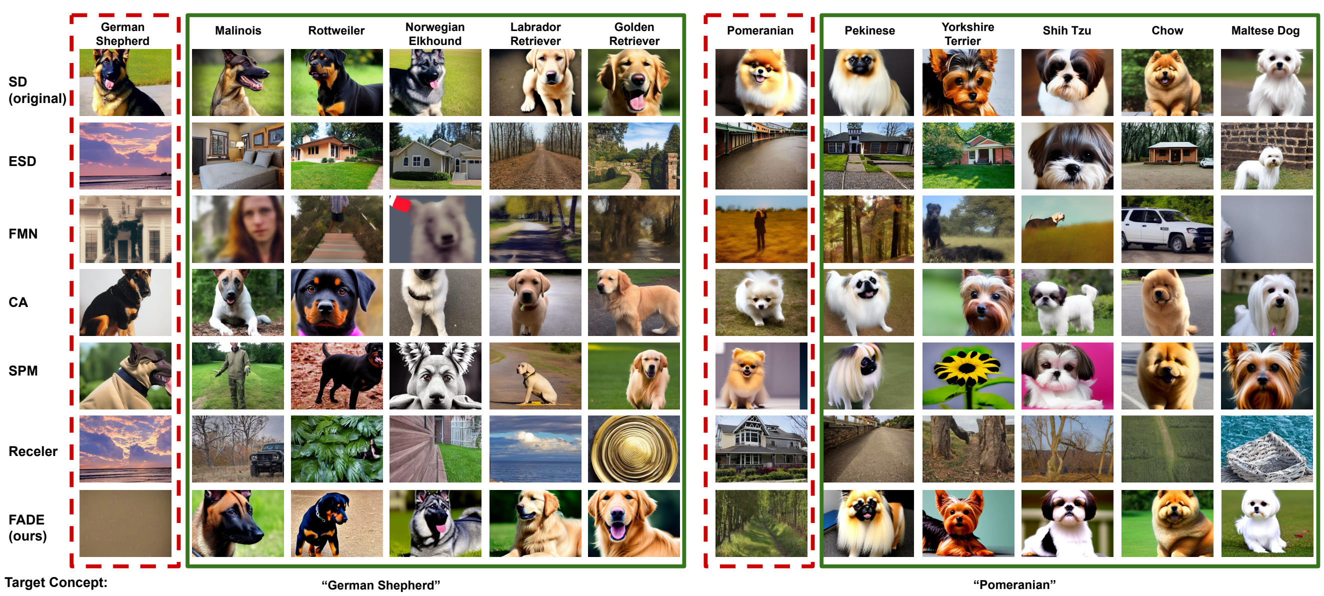

Figure 6. Qualitative comparison of FADE with various algorithms for erasing German Shepherd and Pomeranian while retaining closely looking breeds extracted through concept lattice from the Stanford Dogs dataset.

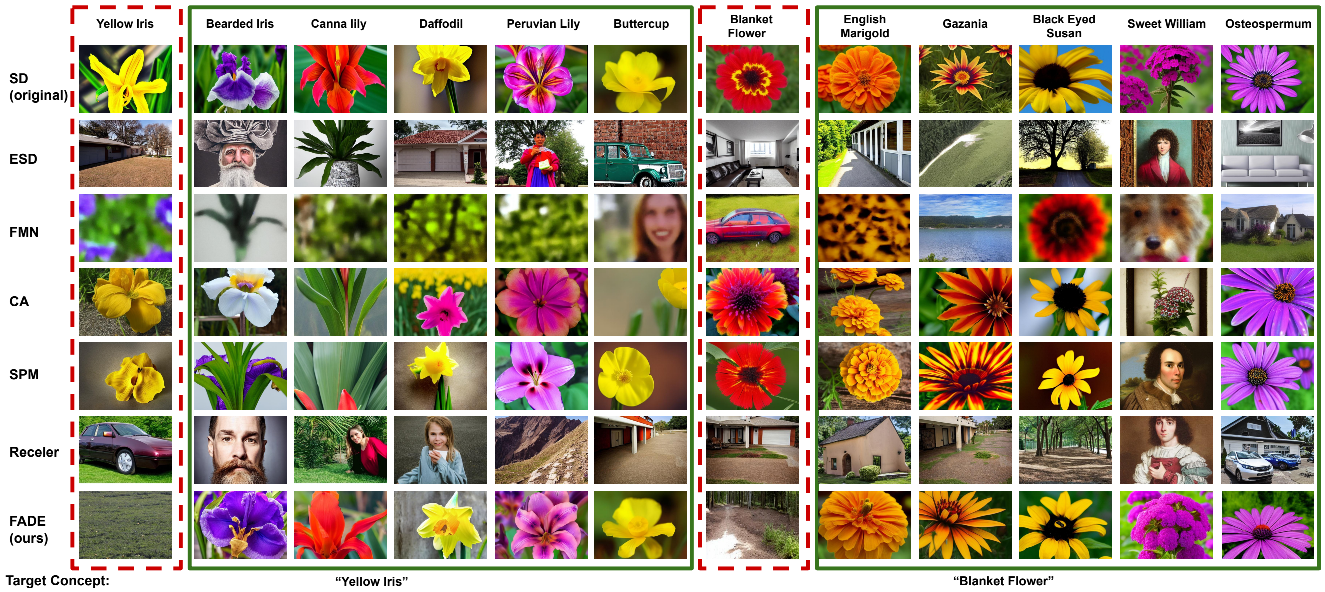

Figure 7. Qualitative comparison of FADE with various algorithms for erasing Yellow Iris and Blanket Flower while retaining other similar-looking flowers through concept lattice from the Oxford Flowers dataset.

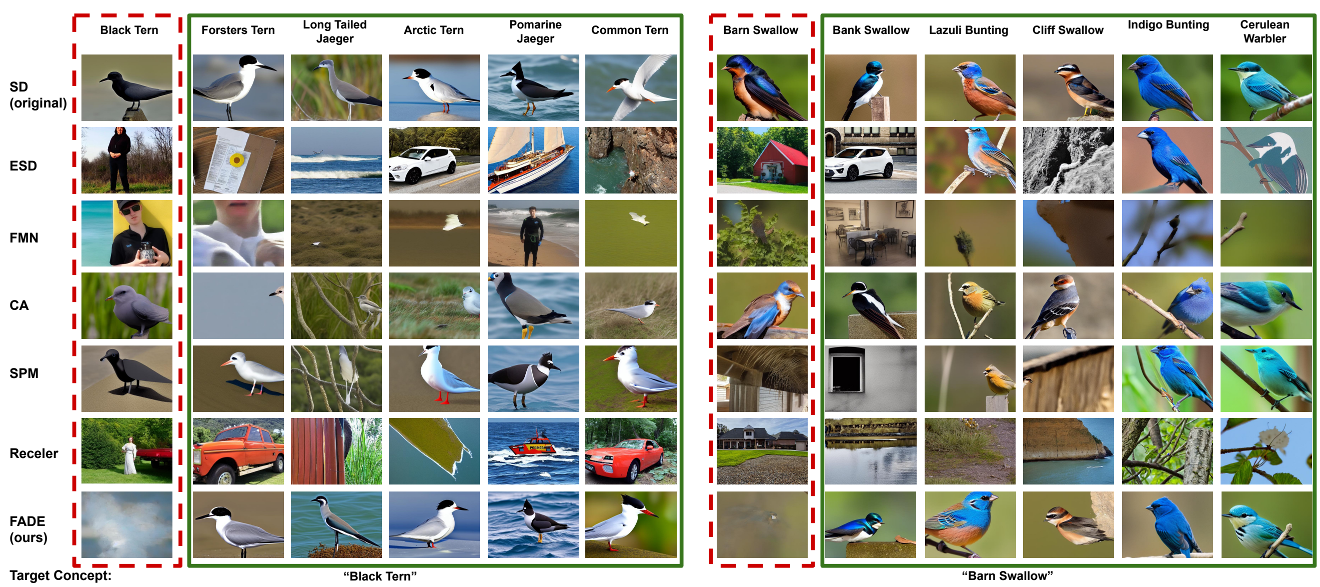

Figure 8. Qualitative comparison of FADE with various algorithms for erasing Blank Tern and Barn Swallow while retaining other closely looking bird species extracted through concept lattice from CUB dataset.

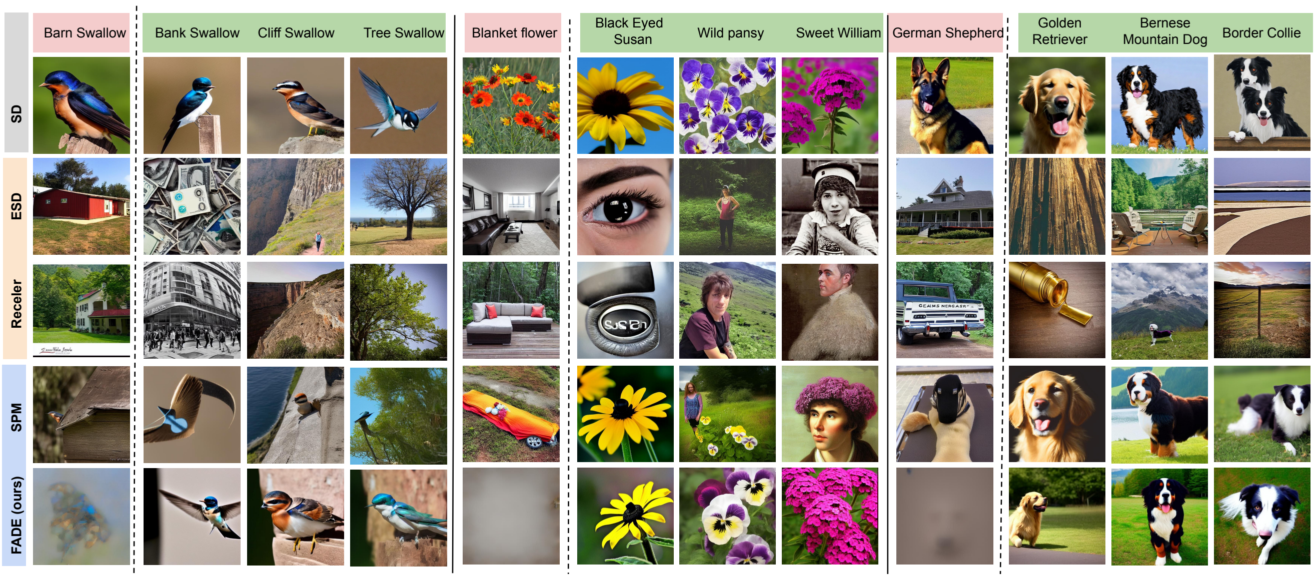

Figure 9. Illustration of concept redirection observed after unlearning target concepts using various algorithms. For ESD and Receler, the erasure of "Blanket Flower" redirects to unrelated outputs, such as a "girl with a black eye" for "Black-eyed Susan flower" and "a man named William" for "Sweet William flower." Similar redirection is seen with bird classes like "Cliff Swallow" and "Tree Swallow." In contrast, SPM and FADE effectively erase target concepts without inducing semantic redirection, ensuring coherence and retention of related knowledge.