

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

Kartik Thakral¹, Tamar Glaser², Tal Hassner³, Mayank Vatsa¹, Richa Singh¹

¹IIT Jodhpur, ²Harman International, ³Weir AI

{thakral.1, mvatsa, richa}@iitj.ac.in, {tamarglasr, talhassner}@gmail.com



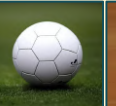
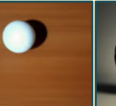







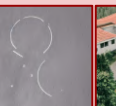






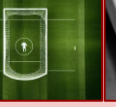
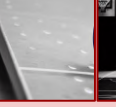

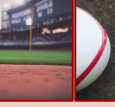
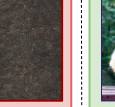


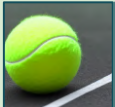

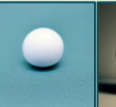




	Target Concept to Unlearn	Fine-Grained Concepts						Coarse-Grained Concepts	Objective
Target Model	Golf Ball	Tennis ball	Soccer ball	Ping-pong ball	Basketball	Baseball	Rugby ball	Corgi	Unlearn Concept
SD v1.4									Preserve Fine-Grained Concepts
Existing Unlearning Methods									
ESD Gandikota et al.									Coarse-Grained Evaluation ✓
SPM Lyu et al.									Fine-Grained Evaluation ✗
FADE (ours)									Coarse-Grained Evaluation ✓ Fine-Grained Evaluation ✓

Figure 1. Fine-Grained Concept Erasure: This figure demonstrates the issue of collateral forgetting (termed as *adjacency*) in selective concept erasure using existing state-of-the-art algorithms in text-to-image diffusion-based foundation models. It highlights the inability of methods that can precisely erase target concepts from a model’s knowledge while preserving its ability to generate closely related concepts.

Abstract

Existing unlearning algorithms in text-to-image generative models often fail to preserve the knowledge of semantically related concepts when removing specific target concepts—a challenge known as *adjacency*. To address this, we propose **FADE** (Fine-grained Attenuation for Diffusion Erasure), introducing adjacency-aware unlearning in diffusion models. FADE comprises two components: (1) the **Concept Neighborhood**, which identifies an adjacency set of related concepts, and (2) **Mesh Modules**, employing a structured combination of Expungement, Adjacency, and Guidance loss components. These enable precise erasure of target concepts while preserving fidelity across related and unrelated concepts. Evaluated on datasets like Stanford Dogs, Oxford Flowers, CUB, I2P, Imagenette, and ImageNet-1k, FADE effectively removes target concepts with minimal impact on correlated concepts, achieving at least a **12% improvement in retention performance** over state-of-the-art methods. Our code and models are available on the project page: [iab-rubric/unlearning/FG-Un](https://iab-rubric.unlearning/FG-Un).

1. Introduction

Text-to-image diffusion models [14, 17, 18] have achieved remarkable success in high-fidelity image generation, demonstrating adaptability across both creative and industrial applications. Trained on expansive datasets like LAION-5B [21], these models capture a broad spectrum of concepts, encompassing diverse objects, styles, and scenes. However, their comprehensive training introduces ethical and regulatory challenges, as these models often retain detailed representations of sensitive or inappropriate content. Thus, there is a growing need for selective concept erasure that avoids extensive retraining, as retraining remains computationally prohibitive [4, 12, 23].

Current generative unlearning methods aim to remove specific concepts while preserving generation capabilities for unrelated classes, focusing on the concept of **locality** [4, 10, 12, 13]. However, these methods often lack fine-grained control, inadvertently affecting semantically similar classes when erasing a target concept (refer Figure 1). This creates the need for **adjacency-aware unlearn-**

ing—the ability to retain knowledge of classes closely related to the erased concept. Specifically, adjacency-aware unlearning seeks to modify a model such that the probability of generating the target concept c_{tar} given input x approaches zero, i.e., $P_{\theta}(c_{\text{tar}}|x) \rightarrow 0$, while ensuring $P_{\theta}(\mathcal{A}(c_{\text{tar}})|x) \approx P_{\theta_{\text{original}}}(\mathcal{A}(c_{\text{tar}})|x)$, where $\mathcal{A}(c_{\text{tar}})$ represents a carefully constructed set of semantically related classes that should remain unaffected by unlearning.

To address these challenges, we introduce **FADE** (*Fine-grained Attenuation for Diffusion Erasure*), a framework for adjacency-aware unlearning in text-to-image diffusion models. FADE has two core components: the **Concept Neighborhood**, which identifies semantically related classes to form an adjacency set using fine-grained semantic similarity, and the **Mesh Modules**, which balance target concept erasure with adjacent class retention through Expungement, Adjacency, and Guidance loss components. This design ensures effective unlearning of target concepts while preserving the integrity of neighboring and unrelated concepts in the semantic manifold. We evaluate FADE using the Erasing-Retention Balance Score (ERB), the proposed metric that quantifies both forgetting and adjacency retention. Experimental results across fine- and coarse-grained datasets—including Stanford Dogs, Oxford Flowers, CUB, I2P, Imagenette, and ImageNet-1k—demonstrate FADE’s effectiveness in erasing targeted concepts while protecting representations of adjacent classes. The key contributions include (i) formalization of adjacency-aware unlearning for text-to-image diffusion models, emphasizing the need for precise retention control, (ii) introduction of FADE, a novel method for unlearning target concepts with effective adjacency retention, and (iii) proposal of the Erasing-Retention Balance Score (ERB) metric, designed to capture both forgetting efficacy and adjacency retention. Using ERB, extensive evaluations are performed on fine- and coarse-grained protocols to assess the erasing performance of FADE compared to state-of-the-art methods.

2. Related Work

Advancements in generative modeling have highlighted the need for effective unlearning techniques. Text-to-image generative models trained on large datasets often encapsulate undesired or inappropriate content, necessitating methods that can selectively remove targeted concepts while preserving overall model functionality. Generative machine unlearning aims to facilitate precise modifications without affecting unrelated knowledge.

Generative Machine Unlearning: Existing approaches focus on unlearning specific concepts from generative models. Gandikota et al. [4] used negative guidance in diffusion models to steer the generation away from unwanted visual elements like styles or object classes. FMN [28] adjusts cross-attention mechanisms to reduce emphasis on

undesired concepts, while Kumari et al. [12] aligned target concepts with surrogate embeddings to guide models away from undesirable outputs. Thakral et al. [24] proposed a robust method of continual unlearning for sequential erasure of concepts. For GANs, Tiwari et al. [25] introduced adaptive retraining to selectively erase classes. However, the high computational cost of retraining remains a drawback [4, 12, 23], highlighting the need for efficient methods.

Parameter-Efficient Fine-Tuning (PEFT) methods address these computational challenges by modifying a small subset of parameters. UCE [5] offers a closed-form editing approach that aligns target embeddings with surrogates, enabling concept erasure while preserving unrelated knowledge. SPM [13] introduced “Membranes,” lightweight adapters that selectively erase concepts by altering model sensitivity. Similarly, Receler [10] incorporates “Erasers” into diffusion models, for robust and adversarially resilient concept erasure with minimal impact on unrelated content.

Fine-Grained Classification: Fine-grained classification tackles the challenge of distinguishing highly similar classes, often complicated by subtle visual differences and label ambiguities. In datasets like ImageNet [19], overlapping characteristics between classes hinder classification accuracy. Beyer et al. [3] and Shankar et al. [22] introduced multi-label evaluation protocols to accommodate multiple entities within a single image, benefiting tasks such as organism classification.

Recent methods have advanced the evaluation of fine-grained errors by automating their categorization. Vasudevan et al. [26] proposed an error taxonomy to distinguish fine-grained misclassifications from out-of-vocabulary (OOV) errors, enabling nuanced analyses of visually similar classes. Peychev et al. [16] automated error classification, providing deeper insights into model behavior in fine-grained scenarios.

Challenges in Adjacency-Aware Erasure: Despite progress, achieving adjacency-aware erasure while maintaining locality remains a significant challenge. Current unlearning methods often struggle with fine-grained concept forgetting, inadvertently affecting the semantic neighborhood of the target concept. This underscores the need for techniques that can selectively remove only the target concept while preserving the integrity of related classes.

3. Fine-Grained Unlearning

3.1. Preliminary

Text-to-image diffusion models have become essential for high-quality image synthesis by learning to generate images through a denoising process [1, 7]. Starting with Gaussian noise, these models refine an image over T timesteps by predicting the noise component $\epsilon_{\theta}(x_t, c, t)$ at each step t , conditioned on a textual prompt c . This reverse process, modeled as a Markov chain, aims to recover the fi-

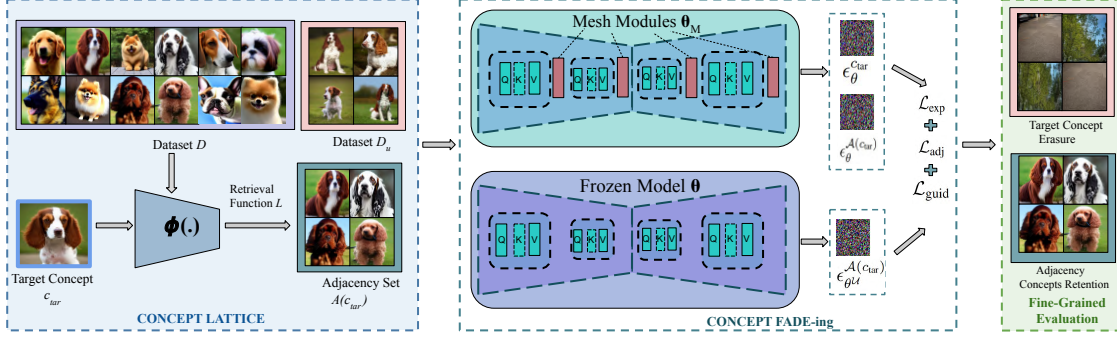


Figure 2. Visual illustration of complete erasure process. (a) The dataset D is organized into unlearning set \mathcal{D}_u and adjacency set $\mathcal{A}(c_{\text{tar}})$ using concept neighborhood, (b) these sets are utilized by mesh-modules for selective erasure while maintaining semantic integrity if the model on neighboring concepts.

nal image x_0 from initial noise x_T , with generation probability defined as $P_\theta(x_{0:T}) = P(x_T) \prod_{t=1}^T P_\theta(x_{t-1}|x_t)$, where $P(x_T)$ is the Gaussian prior. Latent Diffusion Models (LDMs) [18] further improve efficiency by operating in a compressed latent space z , where $z = \mathcal{E}(x)$, and noise is progressively added to obtain z_t . The model learns to minimize the difference between true noise ϵ and predicted noise ϵ_θ , with a training objective:

$$\mathcal{L} = \mathbb{E}_{z,t,c,\epsilon} \left[\|\epsilon - \epsilon_\theta(z_t, c, t)\|_2^2 \right], \quad \epsilon \sim \mathcal{N}(0, 1), \quad (1)$$

Given the high parameter count of these models, efficient fine-tuning for unlearning tasks necessitates Parameter-Efficient Fine-Tuning (PEFT) techniques. We employ a LoRA-based method [9], termed **Mesh Modules** throughout this paper, which selectively updates only a subset of model parameters. Specifically, the weight update ∇W for any pretrained weight matrix $W_0 \in \mathbb{R}^{d \times k}$ is decomposed as $\nabla W = BA$, where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, and $r \ll \min(d, k)$. Only the smaller matrices A and B are trained, preserving computational efficiency and limiting the risk of overfitting by keeping W_0 fixed. This adaptation effectively enables precise concept removal while preserving core generative capabilities.

3.2. Problem Formulation

Our objective is to selectively unlearn a target concept $c_{\text{tar}} \in \mathcal{C}$ in a generative model while preserving its performance on semantically similar (adjacent) and unrelated concepts. Let $\mathcal{D} = \{d_1, d_2, \dots, d_N\}$ represent a dataset where each data point d_i is associated with a subset of concepts $\mathcal{C}_{d_i} \subseteq \mathcal{C}$, with \mathcal{C} representing the universal set of all concepts learned by the model. We denote the pre-trained generative model as θ , mapping input prompts $x \in \mathcal{X}$ to images $y \in \mathcal{Y}$, thereby learning the conditional distribution $P_\theta(y | x)$. To achieve unlearning, we aim to update the model parameters from θ to θ^μ via an unlearning function \mathcal{U} , such that

the probability of generating images associated with c_{tar} i.e., $y_{c_{\text{tar}}}$ approaches zero for any input prompt x , expressed as $P_{\theta^\mu}(y_{c_{\text{tar}}} | x) \rightarrow 0, \forall x \in \mathcal{X}$. Simultaneously, we seek to maintain the model’s performance on adjacent concepts and unrelated concepts.

Let $\mathcal{A}(c_{\text{tar}}) \subseteq \mathcal{C}$ denote the adjacency set, containing concepts closely related to c_{tar} . The unlearning objective must satisfy the following:

1. Retention of Adjacent Concepts:

$$P_{\theta^\mu}(y_c | x) \approx P_\theta(y_c | x), \quad \forall c \in \mathcal{A}(c_{\text{tar}}), \forall x \in \mathcal{X}. \quad (2)$$

2. Preservation of Unrelated Concepts:

$$P_{\theta^\mu}(y_c | x) \approx P_\theta(y_c | x), \quad \forall c \in \mathcal{C} \setminus \mathcal{A}(c_{\text{tar}}) \cup \mathcal{A}(c_{\text{tar}}), \forall x \in \mathcal{X}. \quad (3)$$

3.3. FADE: Fine-grained Erasure

We present FADE (Fine-grained Attenuation for Diffusion Erasure), a method for targeted unlearning in text-to-image generative models, designed to remove specific concepts while preserving fidelity on adjacent and unrelated concepts (see Figure 1). FADE organizes model knowledge into three subsets: the Unlearning Set \mathcal{D}_u , the Adjacency Set \mathcal{D}_a , and the Retain Set \mathcal{D}_r .

The Unlearning Set \mathcal{D}_u consists of images generated using the target concept c_{tar} , such as “Golden Retriever” for a retriever breed class. The Adjacency Set \mathcal{D}_a contains images of concepts similar to c_{tar} (e.g., related retriever breeds), ensuring the erasure of c_{tar} does not compromise the model’s ability to generate closely related classes. We construct \mathcal{D}_a using Concept Neighborhood, which systematically identifies semantically proximal classes to c_{tar} based on similarity scores.

The Retain Set \mathcal{D}_r , containing images of diverse and unrelated concepts (e.g., “Cat” or “Car”), serves as a check for broader generalization retention. While successful retention on \mathcal{D}_a typically implies generalization to \mathcal{D}_r , testing with \mathcal{D}_r ensures no unintended degradation in unrelated areas.

FADE employs a structured mesh to modulate the likelihood of generating images including c_{tar} , gradually attenuating the concept’s influence while preserving related and unrelated knowledge. We formalize this data organization by ensuring $\mathcal{D}_u \cup \mathcal{D}_a \cup \mathcal{D}_r \subseteq \mathcal{D}$, with $\mathcal{D}_u \cap \mathcal{D}_a \cap \mathcal{D}_r = \emptyset$. The complete framework can be visualized in Figure 2.

Concept Neighborhood - Synthesizing Adjacency Set Evaluating unlearning on fine-grained concepts requires an adjacency set \mathcal{D}_a , designed to preserve the model’s performance on concepts neighboring the target concept c_{tar} . Ideally, $\mathcal{D}_a = \{c \in \mathcal{C} \mid \text{sim}(c, c_{\text{tar}}) > \tau\}$, where $\text{sim}(c, c_{\text{tar}})$ represents a semantic similarity function and τ is a threshold for high similarity. However, in the absence of taxonomical hierarchies or semantic annotations (e.g., WordNet synsets), constructing \mathcal{D}_a becomes challenging. To address this, we propose an approximation $\mathcal{A}(c_{\text{tar}})$, termed the *Concept Neighborhood*, which leverages semantic similarities to identify the top-K classes most similar to c_{tar} and thus serves as a practical substitute for \mathcal{D}_a .

To construct $\mathcal{A}(c_{\text{tar}})$, we proceed as follows: for each concept $c \in \mathcal{C}$, including c_{tar} , we generate a set of images $\mathcal{I}_c = \{x_1^c, x_2^c, \dots, x_m^c\}$ using θ , where m is the number of images per concept. Using a pre-trained image encoder $\phi : X \rightarrow \mathbb{R}^d$, we compute embeddings for each image: $\mathbf{f}_i^c = \phi(x_i^c)$ for all $x_i^c \in \mathcal{I}_c$. For each concept c , we then compute the mean feature vector $\bar{\mathbf{f}}^c = \frac{1}{N} \sum_{i=1}^N \mathbf{f}_i^c$ and quantify the semantic similarity between the target concept c_{tar} and every other concept $c \in \mathcal{C} \setminus \{c_{\text{tar}}\}$ by calculating the cosine similarity between their mean feature vectors:

$$L(c_{\text{tar}}, c) = \frac{\langle \bar{\mathbf{f}}^{c_{\text{tar}}}, \bar{\mathbf{f}}^c \rangle}{\|\bar{\mathbf{f}}^{c_{\text{tar}}}\| \|\bar{\mathbf{f}}^c\|}, \quad (4)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, and $\|\cdot\|$ denotes the Euclidean norm. We select the top- K concepts with the highest similarity to c_{tar} to form the adjacency set $\mathcal{A}(c_{\text{tar}}) = \{c^{(1)}, c^{(2)}, \dots, c^{(K)}\}$, where $L(c_{\text{tar}}, c^{(i)}) \geq L(c_{\text{tar}}, c^{(i+1)})$ for $i = 1, \dots, K-1$, and $c^{(i)} \in \mathcal{C} \setminus \{c_{\text{tar}}\}$.

This approach effectively constructs $\mathcal{A}(c_{\text{tar}})$ by capturing the fine-grained semantic relationships inherent in the latent feature space, approximating the ideal \mathcal{D}_a with a data-driven methodology that leverages embedding similarity.

Our Concept Neighborhood method is further supported by a theoretical link between k-Nearest Neighbors (k-NN) classification in latent feature space and the optimal Naive Bayes classifier under certain conditions, as established in the following theorem:

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 $\phi : \mathbb{R}^{h \times w \times c} \rightarrow \mathbb{R}^d$ 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 $N \rightarrow \infty$, the number of neighbors $k \rightarrow \infty$, and $k/N \rightarrow 0$. Specifically,

$$\lim_{N \rightarrow \infty} P(C_{k\text{-NN}}(\phi(\mathbf{x})) = C_{\text{NB}}(\mathbf{x})) = 1. \quad (5)$$

A detailed proof is available in the supplementary material, but intuitively, this result shows that the k-NN classifier in latent space approximates the optimal classifier, supporting the use of feature similarity (via k-NN) to identify semantically similar concepts. Thus, the Concept Neighborhood method approximates the latent space’s underlying semantic structure, effectively constructing $\mathcal{A}(c_{\text{tar}})$ for adjacency preservation in the unlearning framework.

Concept FADE-ing The proposed FADE (Fine-grained Attenuation for Diffusion Erasure) algorithm selectively unlearns a target concept c_{tar} through the mesh M , parameterized by $\theta_M^{\mathcal{U}}$, while maintaining the model’s semantic integrity for neighboring concepts. FADE achieves this by optimizing three distinct loss terms: the Erasing Loss, the Guidance Loss, and the Adjacency Loss.

1. **Erasing Loss (\mathcal{L}_{er}):** This loss is designed to encourage the model to erase c_{tar} by modulating the predicted noise ϵ_{θ} in such a way that the changes in the unlearned model are minimal with respect to semantically related classes in $\mathcal{A}(c_{\text{tar}})$, thereby acting as a regularization term. Concurrently, it drives the model’s representation of the target concept c_{tar} in \mathcal{D}_u , disorienting it away from its initial position. Formally, the Erasing Loss is defined as:

$$\mathcal{L}_{\text{er}} = \max \left(0, \frac{1}{|\mathcal{A}(c_{\text{tar}})|} \sum_{x \in \mathcal{A}(c_{\text{tar}})} \left| \epsilon_{\theta_M^{\mathcal{U}}}^{c_{\text{tar}}} - \epsilon_{\theta}^x \right|_2^2 - \frac{1}{|\mathcal{D}_u|} \sum_{x \in \mathcal{D}_u} \left| \epsilon_{\theta_M^{\mathcal{U}}}^{c_{\text{tar}}} - \epsilon_{\theta}^x \right|_2^2 + \delta \right) \quad (6)$$

where $\epsilon_{\theta}^{c_{\text{tar}}}$ represents the predicted noise for the target concept, ϵ_{θ}^x denotes the predicted noise for samples x in either the adjacency set $\mathcal{A}(c_{\text{tar}})$ or the unlearning set \mathcal{D}_u , and δ is a margin hyperparameter enforcing a minimum separation between the noise embeddings of c_{tar} and its adjacent concepts.

2. **Guidance Loss ($\mathcal{L}_{\text{guid}}$):** The Guidance Loss directs the noise prediction for c_{tar} toward a surrogate “null” concept, allowing unlearning without requiring a task-specific surrogate. Formally, it is defined as:

$$\mathcal{L}_{\text{guid}} = \left| \epsilon_{\theta_M^{\mathcal{U}}}^{c_{\text{tar}}} - \epsilon_{\theta}^{c_{\text{null}}} \right|_2^2 \quad (7)$$

where $\epsilon_{\theta}^{c_{\text{null}}}$ denotes the predicted noise for a neutral or averaged “null” concept in the original model. By di-

recting c_{tar} toward a null state, this loss effectively nullifies the influence of the target concept, facilitating generalized unlearning that is adaptable across tasks without specific surrogate selection [12, 13].

3. **Adjacency Loss (\mathcal{L}_{adj}):** The Adjacency Loss acts as a regularization term, preserving the embeddings of concepts in the adjacency set $\mathcal{A}(c_{\text{tar}})$ in the updated model M_{θ^u} . It penalizes deviations between the original and updated model’s noise predictions for these adjacent concepts, defined as:

$$\mathcal{L}_{\text{adj}} = \frac{1}{|\mathcal{A}(c_{\text{tar}})|} \sum_{x \in \mathcal{A}(c_{\text{tar}})} \left| \epsilon_{\theta^u}^x - \epsilon_{\theta}^x \right|_2^2 \quad (8)$$

where ϵ_{θ}^x and $\epsilon_{\theta^u}^x$ denote the noise predictions for concept x in the original and updated models, respectively. This loss constrains the modified model to retain the structural relationships among adjacent classes, preserving the feature space of $\mathcal{A}(c_{\text{tar}})$ post-unlearning.

The total loss function for the FADE algorithm is a weighted sum of the three loss terms:

$$\mathcal{L}_{\text{FADE}} = \lambda_{\text{er}} \mathcal{L}_{\text{er}} + \lambda_{\text{adj}} \mathcal{L}_{\text{adj}} + \lambda_{\text{guid}} \mathcal{L}_{\text{guid}} \quad (9)$$

where λ_{er} , λ_{adj} , and λ_{guid} are hyperparameters controlling the relative influence of each loss term.

4. Experimental Details and Analysis

Datasets: We evaluate FADE using two protocols: (a) **Fine-Grained Unlearning (FG-Un)**, which focuses on erasing c_{tar} while preserving generalization on challenging concepts in \mathcal{D}_a , and (b) **Coarse-Grained Unlearning (CG-Un)**, which assesses the model’s ability to retain generalization on concepts in \mathcal{D}_r . For FG-Un, we utilize fine-grained classification datasets, including Stanford Dogs [11], Oxford Flowers [15], Caltech UCSD Birds (CUB) [27], and ImageNet-1k [19], due to their closely related classes. We evaluate FADE on three target classes per fine-grained dataset and four target classes in ImageNet-1k. Adjacency sets for these classes are constructed using the Concept Neighborhood. For CG-Un, we follow standard evaluation protocols [4, 10] for the Imagenette [8] and I2P [20] datasets, where evaluations focus on the target class and other classes, regardless of semantic similarity.

Baselines: We compare FADE with state-of-the-art methods for concept erasure, including Erased Stable Diffusion (ESD) [4], Concept Ablation (CA) [12], Forget-Me-Not (FMN) [28], Semi-Permeable Membrane (SPM) [13], and Receler [10]. Open-source implementations and standard settings are used for all baseline evaluations.

Evaluation Metrics: For FG-Un, we measure *Erasing Accuracy* (A_{er}), which quantifies the degree of target concept erasure (higher values indicate better erasure), and *Adjacency Accuracy* (A_{adj}), which evaluates retention across

Comparison of Erasing Methods: Similarity Scores vs Average Adjacency Accuracy

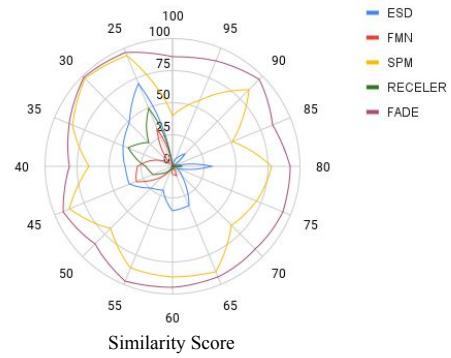


Figure 3. Radar plot comparing FADE with existing unlearning methods (ESD, FMN, SPM, Receler) by structural similarity score (circular axis, %) and adjacency accuracy (radial axes) on concepts from the ImageNet-1k dataset. Most methods begin to degrade beyond a similarity score of 70%, with SPM resilient until 90% and FADE showing the highest robustness. For fair analysis, only methods with $A_{\text{er}} \leq 20\%$ are considered.

$c \in \mathcal{A}(c_{\text{tar}})$. To balance these, we introduce the *Erasing-Retention Balance (ERB) Score*:

$$\text{ERB Score} = \frac{2 \cdot A_{\text{er}} \cdot \hat{A}_{\text{adj}}}{A_{\text{er}} + \hat{A}_{\text{adj}} + \eta}, \quad (10)$$

where $\hat{A}_{\text{adj}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} A_{\text{adj}}$ is the mean Adjacency Accuracy, and η mitigates divide-by-zero errors. The ERB score provides a harmonic mean to evaluate unlearning and retention balance within $\mathcal{A}(c_{\text{tar}})$. For CG-Un, we follow standard protocols for Imagenette and report classification accuracy from a pre-trained ResNet-50 model before and after unlearning. For I2P, we use NudeNet [2] to count nudity classes and FID [6] to measure visual fidelity between the original and unlearned models.

4.1. Results of Fine-Grained Unlearning (FG-Un)

We evaluate FADE’s fine-grained unlearning performance on Stanford Dogs, Oxford Flowers, and CUB datasets, as shown in Table 1. We select three target classes for each dataset and define their adjacency sets using Concept Neighborhood with $K = 5$. To address distribution shifts, we fine-tune pre-trained classifiers on each dataset with samples generated by the SD v1.4 model. We then compute Erasing Accuracy (A_{er}) for the erased target class and Adjacency Accuracy (\hat{A}_{adj}), the mean classification accuracy across adjacency set classes \mathcal{A}_{adj} .

Performance on Fine-Grained Datasets: Table 1 demonstrates that existing algorithms struggle to retain neighboring concepts while erasing the target concept, as reflected by their low ERB scores. FMN shows the weakest adjacency retention, followed by Receler and ESD. CA and

Methods	Metrics	Stanford Dogs			Oxford Flowers			CUB		
		Welsh Springer Spaniel	German Shepherd	Pomeranian	Barbeton Daisy	Yellow Iris	Blanket Flower	Blue Jay	Black Tern	Barn Swallow
ESD [4] (ICCV 2023)	A_{er}	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	\hat{A}_{adj}	20.00	20.00	34.00	48.00	6.00	4.00	15.00	8.00	38.00
	ERB	33.34	33.34	50.74	64.86	11.32	7.69	26.08	14.81	55.07
FMN [28] (CVPRw 2024)	A_{er}	98.80	100.00	98.20	75.60	100.00	63.00	100.00	96.00	98.80
	\hat{A}_{adj}	0.20	0.57	0.60	1.96	7.44	0.84	0.42	2.76	4.28
	ERB	0.39	1.14	1.19	3.82	13.84	1.65	0.84	5.36	8.13
CA [12] (ICCV 2023)	A_{er}	63.00	79.20	68.00	70.20	67.60	27.00	68.60	77.62	42.40
	\hat{A}_{adj}	66.67	63.66	84.40	78.57	55.36	78.60	61.24	54.04	77.92
	ERB	64.75	70.58	75.31	74.15	60.87	40.19	64.71	63.71	54.91
UCE [5] (WACV 2024)	A_{er}	98.20	100.00	100.00	99.00	98.20	99.00	100.00	100.00	100.00
	\hat{A}_{adj}	41.76	46.20	50.72	53.56	39.66	61.24	31.98	34.88	43.44
	ERB	58.80	63.27	67.30	69.51	56.50	75.67	48.46	51.72	60.56
SPM [13] (CVPR 2024)	A_{er}	57.80	99.20	33.60	70.00	48.40	54.00	85.40	86.28	92.60
	\hat{A}_{adj}	65.12	70.80	95.20	91.64	81.68	84.40	80.24	62.16	69.64
	ERB	61.24	82.62	49.66	79.37	60.78	65.86	82.73	72.23	79.49
Receler [10] (ECCV 2024)	A_{er}	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	\hat{A}_{adj}	2.40	2.80	1.16	6.32	0.52	0.88	0.68	0.12	1.28
	ERB	4.68	5.44	2.29	11.88	1.03	1.74	1.35	0.23	2.52
FADE (ours)	A_{er}	99.60	100.00	99.76	99.88	100.00	100.00	100.00	100.00	99.60
	\hat{A}_{adj}	92.60	95.52	94.76	92.44	90.80	91.28	97.28	89.76	95.40
	ERB	95.97	97.70	97.19	96.01	95.17	95.44	98.62	94.60	97.54

Table 1. **Evaluation of erasing breeds of dogs, flowers, and birds from the Stanford Dogs, Oxford Flowers, and CUB datasets, respectively.** A_{er} represents erasing accuracy (higher is better), \hat{A}_{adj} is the mean adjacency set accuracy (higher is better) from concept neighborhood, and ERB reflects the balance between forgetting and retention.

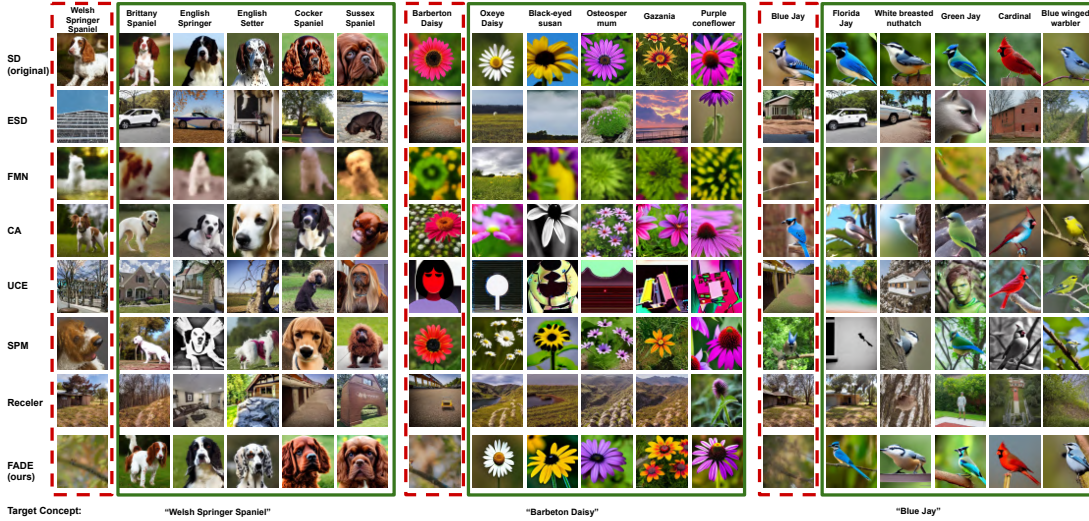


Figure 4. **Qualitative comparison between existing and proposed algorithms for erasing target concepts and testing retention on neighboring fine-grained concepts.** We visualize one target concept each from the Stanford Dogs, Oxford Flowers, and CUB datasets. Visualizations for more concepts are available in the supplementary.

SPM perform moderately, but FADE consistently outperforms all baselines by at least 15% across all target classes. This highlights FADE’s superior ability to balance effective erasure of c_{tar} with the retention of adjacent classes in $\mathcal{A}(c_{tar})$, showcasing its effectiveness in fine-grained unlearning tasks. Further evaluations for top-10 adjacency concepts (available in supplementary) shows the effectiveness of FADE for FG-Un.

Qualitative Analysis: Figure 4 presents generation results for one target class and its adjacency set from each dataset before and after applying unlearning algorithms (additional examples in supplementary material). The first row shows

images generated by the original SD model, followed by results from each unlearning method. Consistent with Table 1, ESD, FMN, and Receler fail to preserve fine-grained details of neighboring classes. CA and SPM retain general structural features but often struggle with specific attributes like color in dog breeds (e.g., Brittany Spaniel, Cocker Spaniel), bird species (e.g., Florida Jay, Cardinal), and flower species. These methods frequently produce incomplete erasure or generalized representations. In contrast, FADE preserves fine-grained details while ensuring effective erasure, as evidenced by sharper distinctions in adjacency sets.

Evaluation on ImageNet-1k Dataset: FADE’s perfor-

	Original SD v1.4		ESD		FMN		CA		SPM		Receler		FADE (ours)	
	$A_{tar} \uparrow$	$A_{others} \uparrow$	$A_{tar} \downarrow$	$A_{others} \uparrow$	$A_{tar} \downarrow$	$A_{others} \uparrow$	$A_{tar} \downarrow$	$A_{others} \uparrow$	$A_{tar} \downarrow$	$A_{others} \uparrow$	$A_{tar} \downarrow$	$A_{others} \uparrow$	$A_{tar} \downarrow$	$A_{others} \uparrow$
Cassette Player	25.00	87.58	0.60	65.50	4.00	20.93	20.20	85.35	2.00	87.31	0.00	77.08	0.00	86.28
Chain Saw	64.00	90.52	0.00	66.66	0.00	39.22	72.80	86.35	20.22	81.44	0.00	70.22	0.00	88.90
Church	82.00	88.27	0.10	69.88	4.00	52.73	47.00	83.64	78.0	87.15	0.80	72.93	0.00	85.15
French Horn	99.8	88.55	0.20	60.55	3.00	38.13	100.00	86.11	13.89	76.91	0.00	66.37	0.00	87.22
Gas Pump	81.85	89.7	4.0	62.71	0.87	39.97	90.60	86.67	16.00	80.26	0.00	66.57	0.00	89.30
Parachute	97.24	86.37	4.0	72.67	11.60	54.33	94.39	85.88	53.60	82.55	1.00	72.57	0.72	84.05
Tench	72.00	88.23	0.00	72.22	1.79	56.22	68.80	86.26	21.80	81.46	3.00	7.66	0.00	87.85
English Springer	97.00	86.40	5.2	68.57	9.40	65.57	35.50	79.11	38.88	81.11	47.88	76.26	0.00	83.75
Garbage Truck	94.64	89.525	0.80	62.57	0.27	21.26	75.00	83.86	27.20	79.42	0.00	64.97	0.00	88.20
Golf Ball	99.85	86.05	0.60	61.50	24.00	54.80	99.60	86.08	50.00	81.75	0.00	69.82	1.6	85.25

Table 2. **Comparative evaluation for coarse-grained unlearning on the Imagenette dataset with existing state-of-the-art methods.** For all models except the original SD model, a lower A_{tar} indicates better erasure of the target concept, and a higher A_{others} represents better retention, as it is the average accuracy on the non-targeted concepts. Except for ‘Cassette Player,’ \hat{A}_{others} is computed over 8 classes, excluding it due to its lower original accuracy for consistency with prior work.

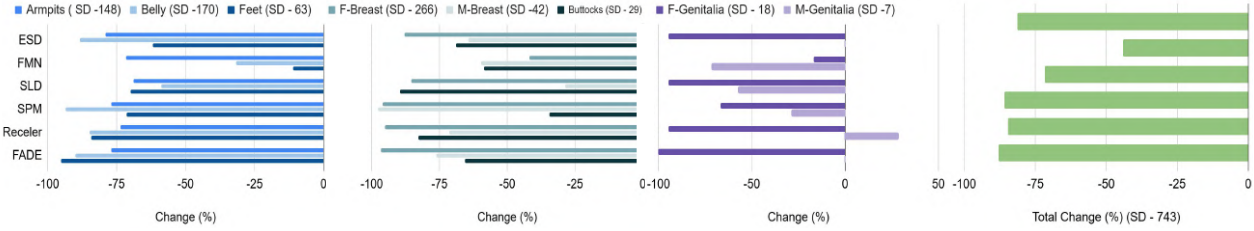


Figure 5. **NudeNet Evaluation on the I2P benchmark.** The numbers followed by ‘SD’ indicate the count of exposed body parts in the SD v1.4 generations. The binplots show the reduction achieved by different methods for erasing nudity. Compared to prior works, FADE effectively eliminates explicit content across various nude categories.

Target Class	Golf Ball	Garbage Truck	English Springer	Tench
ESD	44.81	44.91	50.07	74.74
FMN	49.62	3.30	1.42	56.96
CA	0.79	39.49	63.56	35.72
SPM	63.64	75.18	86.02	72.30
Receler	20.07	32.77	47.62	56.36
FADE (ours)	96.82	91.65	97.93	87.08

Table 3. **Evaluation of erasing structurally similar concepts from ImageNet-1k dataset.** We present the ERB scores, with FADE significantly outperforming all existing algorithms. A_{er} and A_{adj} are available in the supplementary.

mance on ImageNet-1k is evaluated for target classes such as Balls, Trucks, Dogs, and Fish. Using Concept Neighborhood, adjacency sets closely align with the manually curated fine-grained class structure by Peychev et al. [16], demonstrating Concept Neighborhood’s accuracy. Table 2 shows that FADE outperforms all baselines, achieving at least 12% higher ERB scores than SPM, the next-best method. FMN and CA perform poorly in both adjacency retention and erasure. Additional details on adjacency composition and metrics are provided in the supplementary.

Adjacency Inflection Analysis: We evaluate the robustness of algorithms as semantic similarity increases using fine-grained classes from ImageNet-1k and fine-grained datasets. Figure 3 illustrates the relationship between CLIP-based structural similarity (circular axis, %) and average adjacency accuracy (radial axes). FMN and ESD degrade at

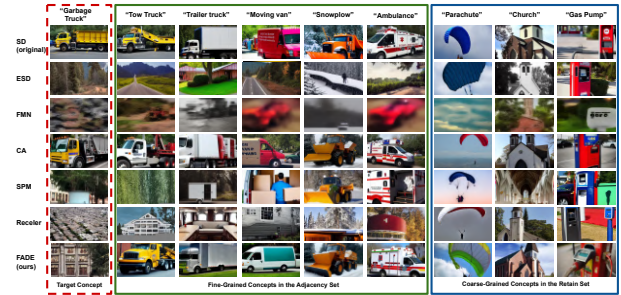


Figure 6. Comparison of FADE with various algorithms for erasing the ‘garbage truck’ class in Fine-Grained and Coarse-Grained Unlearning. The target class, adjacency set and the retain set and constructed from the ImageNet-1k dataset.

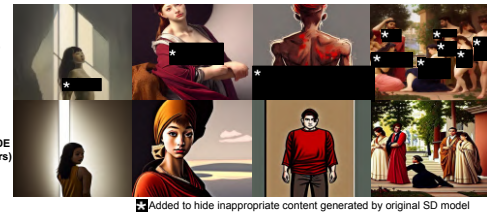


Figure 7. Visualization of before and after unlearning nudity through FADE. The prompts are borrowed from I2P dataset.

78% similarity, with Receler failing at 80%. SPM shows moderate resilience but struggles beyond 90% similarity. In contrast, FADE maintains high adjacency accuracy, demon-

Components			Metrics		
$\mathcal{L}_{\text{guid}}$	\mathcal{L}_{er}	\mathcal{L}_{adj}	$A_{\text{er}} \uparrow$	$\hat{A}_{\text{adj}} \uparrow$	ERB \uparrow
✓	✓	✓	99.60	92.60	95.97
✓	✓	×	25.40	80.24	38.58
✓	×	✓	28.00	95.08	43.26
✓	×	×	31.80	78.16	45.20
×	✓	✓	100.0	76.12	86.44
×	✓	×	43.60	90.44	58.83

Table 4. Ablation study with different components of FADE with target class as Welsh Springer Spaniel.

strating robustness even at high similarity levels, validating its effectiveness in adjacency-aware unlearning tasks.

4.2. Coarse-Grained Unlearning(CG-Un) Results

We evaluate FADE and state-of-the-art methods on the Imagenette dataset, which exhibits minimal semantic overlap. Results are presented in Table 2. For each target class, we report the target erasure accuracy (A_{tar} , lower is better) and the average accuracy on other classes (\hat{A}_{others} , higher is better). These metrics assess erasure on \mathcal{D}_u and retention on \mathcal{D}_r . FADE achieves the best balance between erasure and retention, outperforming all baselines. CA and SPM perform moderately well due to their partial target class removal, which preserves structure and enhances retention. Receler, ESD, and FMN exhibit sub-optimal performance, with FMN being the weakest.

Qualitative Analysis: Figure 4 illustrates qualitative results for the overlapping class of “Garbage Truck” from ImageNet-1k and Imagenette. While ESD, FMN, SPM, and Receler unlearn the target class, they struggle with generalizability across adjacent classes. FADE, in contrast, demonstrates robust generalizability in both FG-Un (ImageNet-1k) and CG-Un (Imagenette), achieving the highest overall performance (Table 2). Additional visualizations for FG-Un and CG-Un classes are included in the supplementary.

Nudity Erasure on I2P: We further evaluate FADE on I2P nudity prompts using NudeNet to detect targeted nudity classes. FADE achieves the highest erasure ratio change of 87.88% compared to the baseline SD v1.4 model, outperforming all methods. Among competitors, SPM ranks second, followed by Receler and ESD. On the nudity-free COCO30K dataset, FADE scores an FID of 13.86, slightly behind FMN (13.52). However, FMN’s erasure ratio change is significantly lower at 44.2%, highlighting its ineffectiveness in nudity erasure. Figure 7 shows qualitative results, illustrating FADE’s superior performance in removing nudity across various prompts.

4.3. Ablation Study

We study the individual contributions of FADE’s loss components: guidance loss ($\mathcal{L}_{\text{guid}}$), erasing loss (\mathcal{L}_{er}), and adjacency loss (\mathcal{L}_{adj}). Table 4 shows results for the target class “Welsh Springer Spaniel” using Erasure Accuracy (A_{er}),

Adjacency Accuracy (\hat{A}_{adj}), and the ERB score. The complete model achieves the highest ERB score of 95.97, balancing target erasure and adjacency preservation. Excluding \mathcal{L}_{adj} results in a sharp drop to 38.58 ERB, highlighting its role in adjacency retention. Removing \mathcal{L}_{er} reduces ERB to 43.26, emphasizing its importance in precise erasure. Similarly, omitting $\mathcal{L}_{\text{guid}}$ achieves perfect erasure accuracy (100.0) but lowers ERB to 86.44, reflecting its necessity for maintaining structural integrity in the adjacency set.

4.4. Qualitative and User Study

We conducted a user study with 40 participants aged 18–89 to evaluate FADE’s performance from a human perspective. Participants assessed both erasure and retention tasks across nine target concepts (see Table 1), each paired with their top three related concepts from the Stanford Dogs, Oxford Flowers, and CUB datasets. For the *erasure evaluation*, participants judged whether images generated by the unlearned models effectively removed the target concept. For the *retention evaluation*, they assessed if adjacent classes were correctly retained. Real examples were provided beforehand to ensure consistency. Each participant evaluated 81 images. Scores from the erasure and retention tasks were aggregated to compute the ERB score for each method. The user study results yielded ERB scores as follows: FADE achieved the highest score of 59.49, outperforming CA (49.38), SPM (49.13), FMN (43.07), ESD (38.43), and Receler (0.06). Participants noted that CA often failed to fully remove the target concept, while Receler adversely affected adjacent classes. These findings highlight FADE’s superiority in balancing effective erasure and retention, as perceived by human evaluators.

5. Conclusion

This work introduces *adjacency* in unlearning for text-to-image models, highlighting how semantically similar concepts are disproportionately affected during erasure. Current algorithms rely on feature displacement, which effectively removes target concepts but distorts the semantic manifold, impacting adjacent concepts. Achieving fine-grained unlearning, akin to creating “holes” in the manifold, remains an open challenge. The proposed FADE effectively erases target concepts while preserving adjacent knowledge through the Concept Neighborhood and Mesh modules. FADE advances adjacency-aware unlearning, emphasizing its importance in maintaining model fidelity.

6. Acknowledgement

The authors thank all volunteers in the user study. This research is supported by the IndiaAI mission and Thakral received partial funding through the PMRF Fellowship.

References

- [1] Stable diffusion. <https://huggingface.co/CompVis/stable-diffusion-v-1-4-original>, 2022. Accessed: 2023-11-09.
- [2] P Bedapudi. Nudenet: Neural nets for nudity classification, detection and selective censoring, 2019.
- [3] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xi-aohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.
- [4] 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.
- [5] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024.
- [6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [8] Jeremy Howard and Sylvain Gugger. Fastai: a layered api for deep learning. *Information*, 11(2):108, 2020.
- [9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [10] Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. *arXiv preprint arXiv:2311.17717*, 2023.
- [11] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, 2011.
- [12] 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.
- [13] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7559–7568, 2024.
- [14] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
- [15] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.
- [16] Momchil Peychev, Mark Müller, Marc Fischer, and Martin Vechev. Automated classification of model errors on imagenet. *Advances in Neural Information Processing Systems*, 36:36826–36885, 2023.
- [17] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- [18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
- [20] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023.
- [21] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.
- [22] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pages 8634–8644. PMLR, 2020.
- [23] Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitriy Pyrkov, Sergei Popov, and Artem Babenko. Editable neural networks. *arXiv preprint arXiv:2004.00345*, 2020.
- [24] Kartik Thakral, Tamar Glaser, Tal Hassner, Mayank Vatsa, and Richa Singh. Continual unlearning for foundational text-to-image models without generalization erosion. *arXiv preprint arXiv:2503.13769*, 2025.
- [25] Piyush Tiwary, Atri Guha, Subhodip Panda, et al. Adapt then unlearn: Exploiting parameter space semantics for unlearning in generative adversarial networks. *arXiv preprint arXiv:2309.14054*, 2023.
- [26] Vijay Vasudevan, Benjamin Caine, Raphael Gontijo Lopes, Sara Fridovich-Keil, and Rebecca Roelofs. When does dough become a bagel? analyzing the remaining mistakes on imagenet. *Advances in Neural Information Processing Systems*, 35:6720–6734, 2022.
- [27] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

- [28] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1755–1764, 2024.