LIFT: FEATURE TRANSPORT FOR UTILITY-PRESERVING CLIP UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

As vision-language models such as CLIP are increasingly deployed in real-world applications, model owners may need to remove sensitive or harmful visual concepts without retraining from scratch. Machine unlearning provides a natural framework for this goal: removing targeted information while preserving the model's general utility. However, suppressing a target class is not enough: the target can remain recoverable from visual features with only a few examples, and aggressive unlearning can degrade model performance on ordinary visual recognition tasks, including out-of-distribution (OOD) recognition. We propose LIFT, a Label-Indistinguishable Feature Transport method for utility-preserving CLIP class unlearning. LIFT treats unlearning as controlled feature relocation: instead of pushing target-class features toward arbitrary directions, it transports them toward cluster-specific distributions of real non-forget image features. These anchor distributions provide a natural destination in the original CLIP feature space. Across fine-grained unlearning benchmarks and comparisons with the most recent unlearning methods, LIFT suppresses all target zero-shot accuracy while making the forgotten classes much harder to recover. For instance, LIFT preserves original CLIP's zero-shot transfer ability while achieving a $\mathbf{48}.\mathbf{4}\%$ relative OOD improvement over the prior state-of-the-art methods. We further evaluate a more challenging selective setting in which only a subset of a dataset's classes is forgotten, while the retained classes are semantically related to the forgotten ones and close to them in CLIP's embedding space. On several datasets spanning car manufacturers, food, remote-sensing scenes, textures, and common objects, LIFT reduces forget accuracy to at most 1.2% while preserving strong accuracy on the retained classes, whereas prior state-of-the-art unlearning methods leave the forgotten classes recognizable or substantially degrade the retained ones.

1 INTRODUCTION

Vision-language models such as CLIP (Radford et al., 2021) are widely applied because their image and text encoders form a shared representation that transfers to many downstream datasets through zero-shot classification. This strength also makes class unlearning difficult. In practice, existing unlearning methods can reduce target recognition, but often at the cost of damaging non-forget and out-of-distribution (OOD) performance (Yang et al., 2025; Hatami et al., 2026; Ha et al., 2025). This leads to the main challenge we address: effective unlearning should make the target class hard to recognize and hard to recover, without degrading CLIP's transferable zero-shot representation.

Recent methods make progress on target unlearning, but they often remove or disrupt target features without defining where features should go. SSD (Foster et al., 2024) dampens forget-important parameters, DAMP (Hatami et al., 2026) removes forget-specific directions, and CLIPErase (Yang et al., 2025) targets visual-textual associations. These strategies can lower target accuracy, but arbitrary feature movement can damage CLIP's image-text geometry for zero-shot transfer. This motivates a different view: unlearning should move forget features away from their original labels and toward a natural non-forget destination.

To address this, we propose LIFT, a **L**abel-**I**ndistinguishable **F**eature **T**ransport method for utility-preserving CLIP class unlearning. The key idea is to move forget features toward natural distributions of real non-forget image features produced by the original CLIP model. We first cluster forget images

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

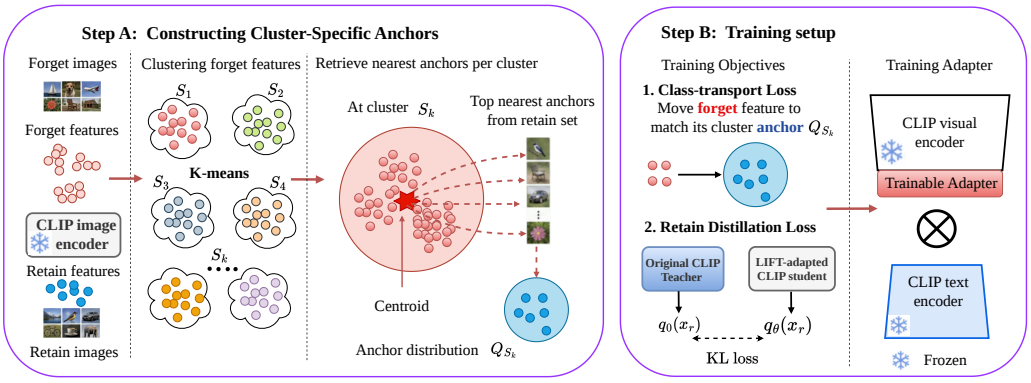


Figure 1: Overview of our method LIFT. We first cluster forget-image features and retrieve cluster-specific non-forget anchors from the retain pool. During training, class transport matches LIFT-transformed forget features to the corresponding anchor distributions, while retain distillation preserves original-CLIP behavior on non-forget images. Only the late visual adapters are trainable; the text encoder and visual-text projection remain frozen.

in the original CLIP feature space and retrieve a cluster-specific set of non-forget anchor images from a retain pool. During training, LIFT matches each group of forget features to its corresponding anchor distribution using a Maximum Mean Discrepancy objective (Gretton et al., 2012).

To preserve the pretrained representation, LIFT updates only a small part of the visual encoder. We insert retain-inactive LoRA adapters into last two visual blocks, keeping the text encoder and most parameters frozen. Unlike standard LoRA (Hu et al., 2022), our adapter fixes its input basis using forget-vs-retain activations and learns only the output factor. This makes the adapter more responsive to forget features and less responsive to retain features, helping preserve OOD zero-shot behavior.

We evaluate LIFT on fine-grained CLIP class unlearning with held-out OOD zero-shot evaluation. In these whole-group experiments, every class of a target dataset is removed. Across target suppression, few-shot recovery, and OOD zero-shot accuracy, LIFT improves the unlearning–utility trade-off over recent state-of-the-art unlearning methods (Cai et al., 2025; Yang et al., 2025; Hatami et al., 2026). In particular, for OOD zero-shot preservation, LIFT achieves dramatically large relative improvements over DAMP (Hatami et al., 2026) of 27.2%, 70.6%, and 54.1% when forgetting Dogs, CUB, and Cars, respectively, and 48.4% on average.

We also evaluate LIFT and other baselines on selective forgetting experiments, where similar classes must remain useful. These challenging experiments which are conducted on additional 6 datasets that span various domains: Food-101 (Bossard et al., 2014), RESISC45 (Cheng et al., 2017), EuroSAT (Helber et al., 2019), DTD (Cimpoi et al., 2014), DomainNet (Peng et al., 2019) and COCO (Lin et al., 2014) measure retained-class precision, local damage to nearby classes, and recovery across several visual domains.

Our contributions are:

- We introduce LIFT, a Label-Indistinguishable Feature Transport method for utility-preserving CLIP class unlearning. LIFT treats unlearning as controlled feature relocation, moving forget features toward real non-forget anchor distributions instead of arbitrary regions of the feature space.
- We design a training objective centered on class transport. The class-transport loss weakens recoverable forget-class geometry by matching forget features to non-forget anchor distributions, while preserving the pretrained visual structure needed for zero-shot transfer.
- We conduct comprehensive experiments on total of nine datasets, evaluate multiple fine-grained forget targets and held-out OOD datasets, showing that LIFT improves the unlearning-utility trade-off over most recent CLIP-unlearning baselines across target zero-shot accuracy, few-shot recovery, and held-out OOD zero-shot evaluation. We extend evaluation to selective requests, cross-domain transfer, sequential requests, and co-occurring objects, and measure both nearby-class retention and resistance to recovery attacks.

2 Related Work

Machine and class unlearning. Machine unlearning studies how to remove the influence of selected data from a trained model without full retraining (Cao & Yang, 2015; Bourtoule et al., 2021). Early work formalizes data deletion and its limitations in learned models (Ginart et al., 2019; Chourasia & Shah, 2023), while recent surveys summarize exact, approximate, privacy-oriented, and verification-focused settings (Wang et al., 2024a). Boundary unlearning shifts decision boundaries to remove target behavior (Chen et al., 2023), Bad Teaching uses an incompetent teacher for class forgetting (Chundawat et al., 2023), SSD (Foster et al., 2024) dampens forget-important parameters, SalUn (Fan et al., 2024) uses gradient-based weight saliency, and DAMP (Hatami et al., 2026) removes forget-specific directions through depth-aware projection. These methods motivate structured forgetting, but most are designed for closed-set classifiers, whereas CLIP unlearning must preserve a shared image–text representation for zero-shot transfer.

Evaluation beyond target accuracy. A recurring issue in unlearning is that low forget accuracy can overstate the degree of removal. Prior work shows that inexact unlearning can give a false sense of privacy (Hayes et al., 2025), and that unlearned models may remain vulnerable to poisoning or relearning effects (Di et al., 2022; Hu et al., 2025; Fan et al., 2025).

Multimodal unlearning. CLIPErase (Yang et al., 2025) directly targets visual-textual associations in CLIP. SLUG (Cai et al., 2025) updates only the single layer whose forget gradient is most important and least aligned with the retain gradient. SAUCE (Li et al., 2025) studies selective concept unlearning in vision-language models with sparse autoencoders, targeted subgroup forgetting studies fine-grained removal inside CLIP classes (Zhang et al., 2025), and MUC (Wang et al., 2024b) studies unlearning for contrastive representation learning with black-box evaluation. LIFT differs by defining an explicit destination for forget features: non-forget anchor distributions in the original CLIP feature space.

Retain preservation and distribution matching. Decoupled distillation (Zhou et al., 2025) separates forgetting and retention objectives, and OOD data can help restore performance after unlearning (Bonato et al., 2024). LIFT also uses non-forget data, but not only as a retention signal: it uses the retain pool to construct anchor distributions that define where forget features should move.

3 Method

3.1 Preliminaries

We study *class unlearning* for CLIP (Radford et al., 2021): given a forget set $\mathcal{D}_f = \{(x_i, y_i)\}$ containing images from the target classes, the goal is to suppress recognition and recoverability of those classes while preserving the model's general zero-shot utility. Let f_0 and g_0 denote the original CLIP visual and text encoders (Radford et al., 2021). We denote the normalized original image feature as $z_0(x) = f_0(x)/\|f_0(x)\|_2$, the normalized text feature as $e(t) = g_0(t)/\|g_0(t)\|_2$, and the normalized image feature after unlearning as $z_\theta(x) = f_\theta(x)/\|f_\theta(x)\|_2$. Throughout LIFT training, the text encoder and visual-text projection remain frozen, and the learnable parameters θ consist only of late visual adapters.

3.2 Constructing Cluster-Specific Anchors

Figure 1 summarizes the pipeline: forget-image features are clustered in the Original CLIP space, each cluster retrieves non-forget anchors from a retain pool, and late visual adapters are trained so that forget features move toward the corresponding anchor distributions while retain behavior is preserved. LIFT constructs the feature distributions that will serve as transport targets for forget images. A single global target would collapse diverse forget images into an unnatural region. LIFT instead builds a separate anchor distribution for each forget-feature cluster. These anchors are real non-forget images from a filtered retain pool, encoded by the original CLIP model.

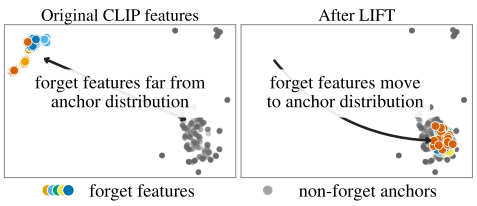


Figure 2: Feature-space visualization of LIFT class transport on CUB dataset (Wah et al., 2011).

We first compute original-CLIP features for the forget training set $\mathcal{D}_f = \{(x_i, y_i)\}$ and the anchor pool $\mathcal{D}_a$. We run K-means directly on the original-CLIP forget features $z_0(x_i)$, assigning each forget image a cluster id $s_i \in \{1, \ldots, K\}$ with $K = 16$. For each cluster s, let m_s be its centroid in the original-CLIP feature space. We retrieve the M nearest anchor images to m_s from the retain anchor pool and denote this set by $\mathcal{A}_s$. The corresponding target distribution is defined as $Q_s = \{z_0(a) : a \in \mathcal{A}_s\}$, i.e., the original-CLIP features of the selected non-forget anchors. Thus, Q_s is a cluster-specific non-forget feature distribution, not a class label and not a single point.

This construction gives each forget cluster a natural landing distribution. The anchor targets are not random vectors, collapsed points, or low-confidence regions: they are original-CLIP feature distributions of real non-forget images. Matching forget features to these distributions, as defined in the next section, keeps LIFT-transformed features close to the visual manifold learned by the original CLIP model while weakening recoverable forget-class geometry. In this way, the anchor construction gives the unlearning update a natural destination and helps preserve OOD zero-shot transfer. Figure 2 illustrates the class-transport mechanism: LIFT moves forget features from their original class-specific regions toward non-forget anchor distributions in the CLIP feature space.

3.3 Learnable Visual Adapter

We first describe which parts of CLIP are trainable during unlearning. LIFT applies adapters only to the last two residual blocks of the CLIP visual encoder, while earlier visual blocks, the final visual-text projection, and the text encoder remain frozen. We call this module a *retain-inactive LoRA* adapter: it keeps the low-rank update form of LoRA (Hu et al., 2022), but fixes the input basis before training so that the adapter is less active on retain features. For any selected linear layer ℓ with frozen weight W_ℓ and input activation h, the adapted output is

$$W_\ell h + \frac{\alpha}{r} A_\ell V_\ell^\top h, \tag{1}$$

where A_ℓ is trainable, while V_ℓ is fixed. In contrast to standard LoRA, which learns both low-rank factors, LIFT only learns the output factor A_ℓ. The fixed basis V_ℓ is estimated from forget-vs-retain layer-input covariances, choosing directions that are more active for forget images and less active for retain images. As a result, the adapter has a natural path for changing forget features, while retain features produce a smaller adapter response. This design lets LIFT suppress target-class evidence while limiting unnecessary changes to non-forget visual representations, helping preserve OOD zero-shot accuracy. We use rank $r = 16$ and scale $\alpha = 16$.

3.4 Overall Training Objective

LIFT uses class transport, $\mathcal{L}_{\text{class-transport}}$, as the main objective, and adds two auxiliary terms: a forget-text regularizer, $\mathcal{L}_{\text{text}}$, and a retain distillation loss, $\mathcal{L}_{\text{retain}}$:

$$\mathcal{L}_{\text{LIFT}} = \mathcal{L}_{\text{class-transport}} + \lambda_1 \mathcal{L}_{\text{text}} + \lambda_2 \mathcal{L}_{\text{retain}} \tag{2}$$

where λ_1 and λ_2 weight the forget-text regularizer and the retain distillation loss, respectively. We show how to determine their values in experiments.

Class-transport loss. After constructing the cluster-specific anchors, each minibatch contains forget images with their original class label c and cluster id s. For a valid pair (c, s), let $X_{c,s}$ be the LIFT-transformed forget features from class c and cluster s, and let Y_s be the sampled original-CLIP anchor features from the corresponding anchor set $\mathcal{A}_s$. We compare these two feature sets with a mixture-of-RBF maximum mean discrepancy (MMD) (Gretton et al., 2012). For normalized features x and y, define the mixture-of-RBF kernel

$$\kappa(x, y) = \frac{1}{|\mathcal{S}|} \sum_{\sigma \in \mathcal{S}} \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2} \right). \tag{3}$$

We use kernel scales $\mathcal{S} = \{0.05, 0.1, 0.2, 0.5, 1.0\}$. For feature sets $X = \{x_i\}_{i=1}^n$ and $Y = \{y_j\}_{j=1}^m$, the biased empirical MMD is

$$\text{MMD}^2(X, Y) = \frac{1}{n^2} \sum_{i,i'} \kappa(x_i, x_{i'}) + \frac{1}{m^2} \sum_{j,j'} \kappa(y_j, y_{j'}) - \frac{2}{nm} \sum_{i,j} \kappa(x_i, y_j). \tag{4}$$

The class-transport loss averages this discrepancy over valid class-cluster pairs:

$$\mathcal{L}_{\text{class-transport}} = \frac{1}{|\mathcal{C}_b|} \sum_{(c,s)\in\mathcal{C}_b} \text{MMD}^2(X_{c,s}, Y_s). \tag{5}$$

Here $\mathcal{C}_b$ contains minibatch pairs with at least two forget samples and at least two sampled anchors. This is a distribution-level loss: it does not force one forget image to match one anchor image. Instead, it makes each class-cluster group of forget features resemble the corresponding anchor distribution. Since anchors from the same Y_s are shared across forget classes inside cluster s, this weakens recoverable class-specific geometry while keeping the transformed features close to non-forget original-CLIP feature regions.

Forget-text regularizer. Class transport moves forget features toward non-forget anchor distributions, but it does not directly constrain similarity to the original forget-class text prompts. We therefore add a simple regularizer between each forget image and its own class text embedding. Let $e_f(y_i)$ be the frozen, normalized original-CLIP text embedding of the forget-class prompt for label y_i, and define the cosine score $r_i = z_\theta(x_i)^\top e_f(y_i)$ for $(x_i, y_i) \in \mathcal{B}_f$. We use

$$\mathcal{L}_{\text{text}} = \frac{1}{|\mathcal{B}_f|} \sum_i \max(0,\ r_i - \tau_{\text{text}}). \tag{6}$$

We set $\tau_{\text{text}} = 0.05$ in all experiments.

Retain distillation loss. For retain images, we keep the model close to original CLIP in the zero-shot prediction space. Let $q_0(x_r)$ be the original-CLIP probability distribution over the retain text bank, and let $q_\theta(x_r)$ be the corresponding LIFT distribution for the same retain image x_r. We use

$$\mathcal{L}_{\text{retain}} = \mathbb{E}_{x_r\in\mathcal{B}_r} \text{KL}(q_0(x_r) \,\|\, q_\theta(x_r)) \,. \tag{7}$$

This loss reduces unnecessary representation shift on non-forget images and helps preserve the visual-text geometry needed for zero-shot transfer.

4 Why Class Transport Helps

Let m_s be the centroid of forget cluster s, and let a be an anchor feature selected for that cluster. Any original forget feature x assigned to s satisfies

$$\|x - a\|_2^2 \le 2\|x - m_s\|_2^2 + 2\|a - m_s\|_2^2. \tag{8}$$

This follows from $x - a = (x - m_s) + (m_s - a)$ and $\|u + v\|_2^2 \le 2\|u\|_2^2 + 2\|v\|_2^2$. K-means reduces the first term on average by forming compact local clusters, while nearest-anchor selection reduces the second term among the available candidates. Together, they encourage low-displacement local transport rather than movement toward one global or arbitrary destination.

Let $X_{c,s}$ and $X_{c',s}$ be the transformed features of two forget classes in cluster s, and let Y_s be their shared anchor feature set. MMD is the distance between empirical kernel mean embeddings (Gretton et al., 2012), so

$$\text{MMD}(X_{c,s}, X_{c',s}) \le \text{MMD}(X_{c,s}, Y_s) + \text{MMD}(X_{c',s}, Y_s). \tag{9}$$

Class transport reduces the squared MMD terms corresponding to the right-hand side; reducing each nonnegative squared discrepancy also reduces its MMD. This lowers an upper bound on the pairwise discrepancy within the cluster.

Importantly, each class-cluster group is matched to a distribution of nearby non-forget anchors, not a single point. Under exact matching, the transformed features follow the variability of Y_s. Local destinations limit unnecessary displacement, while retain distillation constrains changes on non-forget images. Combined with the **local destination** defined by clustering and nearest-anchor selection, this avoids requiring the forget features to move unnecessarily far or toward arbitrary regions of the pretrained feature space. Together, these choices reduce recoverable class structure while limiting unnecessary distortion of the visual representation and helping preserve CLIP's general utility.

Table 1: Whole-group forgetting (%, 3-seed mean; column groups indexed by the forget target; metrics as in Section 5.1, OOD-ZS averaged over the three held-out datasets). Best unlearning result per column in **bold**, second best underlined.

	Target-ZS ↓				OOD-ZS ↑			GDA@4 ↓		
Method	Cars	CUB	Dogs	Avg.	Cars	CUB	Dogs	Cars	CUB	Dogs
Original CLIP (Radford et al., 2021)	65.51	55.09	59.12	59.91	54.33	54.33	54.33	64.50	62.43	44.00
SLUG (Cai et al., 2025)	11.04	3.63	1.60	5.42	37.44	22.04	5.02	60.66	58.94	27.09
CLIPErase (Yang et al., 2025)	0.63	0.58	**0.76**	0.66	7.86	7.06	2.33	11.74	11.34	12.19
DAMP (Hatami et al., 2026)	7.08	6.13	4.36	5.86	35.08	31.35	42.52	28.20	34.18	16.51
LIFT (ours)	**0.46**	**0.53**	**0.76**	**0.58**	**54.06**	**53.49**	**54.09**	**4.72**	**8.20**	**8.15**

5 Whole-Group Forgetting Experiments

5.1 Experimental Setup

We use the same pretrained OpenAI CLIP ViT-B/16 checkpoint (Radford et al., 2021) for LIFT and all baselines, and refer to it as *Original CLIP*. All methods are evaluated with the same forget splits, retain data, and random seeds. All results are reported as the mean over three independent runs. We use $\lambda_1 = 1.0$ and $\lambda_2 = 3.0$ in every experiment. Appendix C (Table 13) conduct detailed ablations on our hyperparameter choices. We train the LIFT adapters with AdamW optimizer (Loshchilov & Hutter, 2017) for 800 steps, $M = 256$ anchors, using learning rate 2×10^{-4}, weight decay 10^{-4} and batch size 128. LIFT has approximately 223K trainable parameters and takes about 15 minutes per run on a single 24GB NVIDIA L4 GPU.

Forget and retain data. We evaluate class unlearning on three fine-grained target datasets: Stanford Dogs (Khosla et al., 2011), CUB (Wah et al., 2011), and Stanford Cars (Krause et al., 2013). For each target, the forget set is the full target training split: 6,122 images for Dogs, 5,994 images for CUB, and 8,144 images for Cars. Caltech-101 (Fei-Fei et al., 2004) is used as the non-forget data source. LIFT uses a fixed 256-image Caltech-101 subset for retain distillation and a larger Caltech-101 pool for anchor retrieval. To avoid semantic leakage, classes with name overlap to the current forget target are removed from the anchor pool.

OOD evaluation. To test whether unlearning preserves CLIP's transferable zero-shot representation, we evaluate clean zero-shot accuracy on three held-out OOD datasets: Oxford Flowers (Nilsback & Zisserman, 2008), FGVC Aircraft (Maji et al., 2013), and UCF-101 (Soomro et al., 2012). In the whole-group experiments, these datasets are never used for forgetting, retain distillation, anchor retrieval, or model selection.

Baselines. We compare LIFT with the most recent unlearning methods: DAMP (Hatami et al., 2026), CLIPErase (Yang et al., 2025), and SLUG (Cai et al., 2025). All baselines start from Original CLIP and use the same forget and retain splits as LIFT whenever the method requires retain data.

Metrics. We report three main metrics. **Target zero-shot accuracy** (Target-ZS) measures zero-shot accuracy on the forget test split. Lower values indicate stronger target suppression. **OOD zero-shot accuracy** (OOD-ZS) measures clean zero-shot accuracy on the held-out OOD utility suite. Higher values indicate better preservation of CLIP's pretrained transfer ability, with Original CLIP as the reference. **Few-shot recovery accuracy** (GDA@k) trains a k-shot Gaussian Discriminant Analysis (GDA) classifier on frozen visual features from the unlearned model and evaluates it on forget-class test images. Lower values indicate that the forgotten classes are harder to recover from the representation.

5.2 Experimental Results

Forget-class zero-shot accuracy. Table 1 reports zero-shot accuracy on the forget test classes after unlearning. LIFT achieves the lowest average target accuracy across Dogs, CUB, and Cars. It matches CLIPErase on Dogs and improves over all baselines on CUB and Cars. DAMP suppresses target recognition, but its accuracy remains higher on some targets, especially on CUB/Cars. SLUG suppresses the targets less consistently: 1.60% on Dogs, but 3.63% on CUB and 11.04% on Cars.

Table 2: Ablating the class-transport distance on whole-group forgetting. Cosine and Mean-square-error (MSE) replace the MMD of LIFT by pointwise matching to the same cluster-selected anchors.

	Target-ZS ↓			OOD-ZS ↑			GDA@4 ↓		
Distance	Cars	CUB	Dogs	Cars	CUB	Dogs	Cars	CUB	Dogs
Cosine	0.45	0.51	0.60	54.34	53.92	54.04	34.60	35.16	34.62
MSE	2.96	0.50	0.62	54.27	54.31	54.11	35.68	34.43	18.59
MMD (LIFT)	0.46	0.53	0.76	54.06	53.49	54.09	4.72	8.20	8.15

OOD zero-shot preservation. Table 1 (per-dataset accuracies in Table 5, Appendix A) tests whether unlearning preserves the transferable zero-shot ability of Original CLIP (Radford et al., 2021) on datasets that are not used for unlearning process. LIFT remains close to Original CLIP across all three forget targets: its average OOD accuracy is 54.09 after forgetting Dogs, 53.49 after forgetting CUB, and 54.06 after forgetting Cars, compared with 54.33 for Original CLIP. DAMP (Hatami et al., 2026) keeps part of the representation useful, but its OOD average drops strongly, especially for CUB and Cars. CLIPErase (Yang et al., 2025) often collapse the zero-shot representation almost completely, despite suppressing the forget target. This shows that low target accuracy alone is not sufficient for CLIP unlearning: a method can forget by damaging the pretrained representation. SLUG (Cai et al., 2025), whose single-layer edit is not constrained by any retain objective on the zero-shot predictions, keeps 37.44% OOD accuracy after forgetting Cars but falls to 22.04% on CUB and 5.02% on Dogs.

Few-shot recovery under GDA. Table 1 (recovery at 1–16 shots in Table 6, Appendix A) evaluates whether the forgotten classes can be recovered from the frozen visual features using a k-shot GDA classifier trained on a few labeled examples per forget class. LIFT achieves the lowest recovery accuracy across all forget targets and all support-shot settings, indicating that the forgotten classes are consistently harder to recover from its visual representation. DAMP (Hatami et al., 2026) reduces recovery compared with Original CLIP, but leaves substantially more recoverable signal, especially at higher shots. CLIPErase (Yang et al., 2025) also lowers recovery, but its OOD zero-shot accuracy collapses in Table 1, suggesting that part of its forgetting comes from damaging the pretrained representation. SLUG leaves the forgotten classes almost fully recoverable: its four-shot recovery is 60.66%, 58.94%, and 27.09% on Cars, CUB, and Dogs, within 5 points of Original CLIP on the first two, so its zero-shot suppression does not reflect a change of the visual features.

5.3 Ablation study

Class-transport matching objective. Table 2 compares our MMD-based class transport with pointwise cosine and MSE matching to cluster-selected anchors. The pointwise variants can reduce target zero-shot accuracy, and in a few cases they are slightly better than MMD on this single metric, but in a very small margin. However, this exposes a serious failure mode for unlearning: although the zero-shot classifier is suppressed, the forget classes remain easily recoverable from the visual features using only four labeled examples per class. For instance, cosine matching leaves 34.62%, 35.16%, and 34.60% recovery on Dogs, CUB, and Cars, respectively, while mean-squared-error matching gives 18.59%, 34.43%, and 35.68%. MMD gives a much more reliable form of forgetting because it matches each group of forget features to an anchor distribution, rather than forcing individual forget images toward individual anchor features. These results support the main design of LIFT: class transport should operate at the distribution level, so that unlearning weakens recoverable feature geometry instead of only lowering target zero-shot accuracy.

Loss terms and adapter. Appendix A (Table 8) ablates the objective and the adapter on whole-group forgetting. Retain distillation alone keeps OOD accuracy but does not forget (average target-ZS 57.15%); class transport alone lowers it to 0.86% and four-shot recovery from 53.90% to 5.17%, the retain term raises average OOD-ZS from 53.58% to 53.93%, and the forget-text regularizer lowers target-ZS further to 0.58% at a small recovery cost. Replacing the retain-inactive adapter by a standard LoRA lowers OOD-ZS on every target, by up to 1.65 points on CUB.

Robustness analysis. Appendix A adds two whole-group studies. Against four recovery attacks (prompt changes, full-support linear probes, retrieval, and LoRA relearning), LIFT keeps linear-probe recovery at 2.96–13.88% against 27.36–36.19% for DAMP, and it needs 49–55 relearning steps where DAMP needs 10–13 (Table 9). CLIPErase resists longer only because its OOD accuracy is

Table 3: Continual unlearning after Dogs→CUB→Cars. Each request unlearned previous classes.

Method	Target-ZS↓	GDA@4↓	OOD↑
Original CLIP (Radford et al., 2021)	59.12 / 55.09 / 65.51	44.0 / 62.4 / 64.5	54.33
SLUG (Cai et al., 2025)	1.15 / 0.53 / 0.65	6.2 / 11.8 / 5.3	2.57
CLIPErase (Yang et al., 2025)	0.76 / 0.49 / 0.45	10.6 / 9.7 / 5.6	1.78
DAMP (Hatami et al., 2026)	3.24 / 3.95 / 3.08	14.9 / 27.5 / 19.2	18.83
LIFT	0.66 / 0.56 / 0.37	8.4 / 8.9 / 4.1	52.53

already 2.33–7.86%. Over $K \in \{4, \ldots, 64\}$ clusters and $M \in \{64, \ldots, 512\}$ anchors, target-ZS stays at or below 1% and OOD accuracy moves by at most 0.30 points (Table 10). Appendix C gives the sensitivity to λ_1 and λ_2, and Appendix E reports the comparisons on a second backbone, CLIP ViT-L/14.

Sequential and joint requests. Table 3 applies three whole-group requests to one model in sequence. LIFT keeps all three targets suppressed and attains 52.53% OOD accuracy, compared with 54.33% before editing. The baselines lose substantially more OOD utility. SLUG ends the chain at 2.57% OOD accuracy, and its first request already leaves only 5.02%, so the later requests have little left to remove. This mimics deployment, where removal requests arrive one after another for the same model. In a separate joint request, LIFT forgets all 516 classes of Dogs, CUB, and Cars together. It reaches 0.40% Target-ZS, 7.37% GDA@4, and 15.29% GDA@full recovery, with 53.71% OOD accuracy.

6 Selective Forgetting Experiments

In this setting, we forget part of a dataset while retaining its neighbor classes that are similar. These requests test more challenging and realistic scenarios to see whether suppression can be localized to a manufacturer, a dish, a scene type, or a visual style while keeping knowledge about their similar neighbor classes.

Experimental setup. We use cosine similarity of Original CLIP training-image prototypes to retrieve three nearest retained classes of each forget class and additionally evaluate the accuracy for these classes which we denote "Near-Retain accuracy". In addition to Cars, Dogs, and CUB, we also conduct our experiments on several datasets spanning different domains, such as: Food-101 (Bossard et al., 2014), RESISC45 (Cheng et al., 2017), EuroSAT (Helber et al., 2019), DTD (Cimpoi et al., 2014), DomainNet (Peng et al., 2019), and COCO (Lin et al., 2014). For Cars, Food-101, EuroSAT, and DTD we use the few-shot dataset splits released by Zhou et al. (2022). Table 4 reports the main comparisons (full versions in Appendix B).

Similar semantic-level forgetting. We forget each of the six car manufacturers with at least ten models and retain all other models. Table 4 (left) shows that LIFT reduces Forget-ZS to 0.0–1.2%, with Retain-ZS within 1.0 point of Original CLIP. Near-Retain stays within 2.2 points for five manufacturers. GDA@4 falls to 13.2-20.3%, while OOD preserves strong accuracy.

DAMP preserves the retained models but leaves substantial recognition and recovery of the requested manufacturer. CLIPErase suppresses zero-shot recognition, but its Retain-ZS falls to 3.4–12.4% and visual recovery remains high. On Food-101, LIFT reduces Forget-ZS from 80.6% to 1.1%, while Retain-ZS and Near-Retain change by only 0.2 and 0.8 points (Table 4, right). Strong suppression also extends to aerial scenes and textures, without a comparable loss of retained-class accuracy. DAMP achieves partial suppression on these domains but generally leaves higher recovery. CLIPErase can have lower recovery, but substantially damages retention and external utility. SLUG suppresses the requested manufacturers (0.3–2.9%) but loses 20–24 points of Retain-ZS and 23–41 points of Near-Retain, and its features stay as recoverable as those of Original CLIP (GDA@4 within 1 point); on the other domains it forgets less (0.9–25.6%) while OOD accuracy drops to 9.3–49.1%.

Component ablations under selective forgetting. Table 14 (Appendix D) extends the design ablations to four selective requests and reports Retain and Near-Retain accuracy. The class-cluster grouping protects the retained classes: replacing it with per-image matching lowers Near-Retain accuracy by 8.8–21.2 points. The MMD objective keeps the forgotten classes hard to recover: pointwise matching raises four-shot recovery by 16–61 points, back to the level of Original CLIP on

Table 4: Selective forgetting (%, 3-seed mean). Left: every Stanford Cars manufacturer with at least ten models. Right: requests across visual domains. Original CLIP (Radford et al., 2021) is the non-unlearned reference. Forget / Retain / Near = zero-shot accuracy on the forgotten classes / all retained in-domain classes / the three nearest retained classes of each forgotten one; GDA@4 = four-shot recovery. DAMP (Hatami et al., 2026) and SLUG (Cai et al., 2025) use all remaining classes of the dataset as retain data, as in their papers; CLIPErase (Yang et al., 2025) uses the retain set of LIFT. Full versions with standard deviations, 16-shot recovery and COCO retrieval are detailed in Tables 11 and 12, Appendix B.

Method	Forget↓	Retain↑	Near↑	GDA@4↓	OOD↑
Audi (14 / 196)					
Original CLIP	55.2	66.2	74.5	48.1	54.3
SLUG	1.5	45.5	40.8	47.4	51.9
CLIPErase	0.0	12.0	7.9	49.2	13.9
DAMP	54.1	65.8	74.4	47.2	54.4
LIFT	0.3	65.9	73.1	15.7	54.1
BMW (13 / 196)					
Original CLIP	62.6	65.6	66.9	51.3	54.3
SLUG	0.3	45.7	43.5	50.6	51.8
CLIPErase	0.0	6.6	2.9	55.4	11.9
DAMP	57.2	65.6	67.1	50.3	53.9
LIFT	0.0	65.3	66.8	13.2	54.2
Chevrolet (22 / 196)					
Original CLIP	47.2	67.7	58.9	61.0	54.3
SLUG	1.0	44.2	30.3	60.2	51.1
CLIPErase	0.0	3.4	2.0	57.9	9.7
DAMP	43.2	67.6	59.4	60.2	53.6
LIFT	1.2	67.5	60.5	16.9	54.1
Dodge (15 / 196)					
Original CLIP	56.6	66.1	62.1	71.4	54.3
SLUG	1.2	44.2	36.9	70.4	52.1
CLIPErase	0.0	6.4	1.2	66.4	15.8
DAMP	52.8	65.7	61.7	69.1	53.8
LIFT	0.3	65.9	59.9	18.0	54.3
Ford (12 / 196)					
Original CLIP	69.7	65.1	44.4	77.2	54.3
SLUG	0.9	42.5	17.2	76.4	51.3
CLIPErase	0.1	12.1	4.3	80.4	25.8
DAMP	65.7	65.3	44.8	76.8	54.1
LIFT	0.2	65.0	45.4	18.7	54.1
Hyundai (11 / 196)					
Original CLIP	45.3	66.6	69.4	75.1	54.3
SLUG	2.9	43.6	28.2	75.1	53.0
CLIPErase	0.0	12.4	3.3	70.6	15.8
DAMP	40.3	66.9	71.0	73.4	54.2
LIFT	0.0	65.6	64.3	20.3	54.2

Method	Forget↓	Retain↑	Near↑	GDA@4↓	OOD↑
Food-101 (Bossard et al., 2014): related dishes (10 / 101)					
Original CLIP	80.6	86.5	85.9	95.5	54.3
SLUG	2.7	25.8	20.8	94.6	28.0
CLIPErase	0.0	1.1	0.0	24.6	0.8
DAMP	67.9	86.8	86.4	94.3	54.1
LIFT	1.1	86.3	85.1	29.4	54.5
RESISC45 (Cheng et al., 2017): related scenes (5 / 45)					
Original CLIP	55.5	62.4	61.6	97.2	54.3
SLUG	22.1	55.2	49.9	96.9	45.5
CLIPErase	6.7	5.9	0.0	63.5	17.3
DAMP	28.1	63.5	62.9	94.0	54.3
LIFT	0.0	62.3	58.8	49.3	54.3
EuroSAT (Helber et al., 2019): land-cover classes (3 / 10)					
Original CLIP	65.5	41.9	49.2	90.4	54.3
SLUG	21.5	34.1	45.5	90.4	49.1
CLIPErase	0.0	14.3	0.0	77.6	20.3
DAMP	22.1	43.1	50.5	80.3	53.9
LIFT	0.4	44.1	51.9	53.7	54.2
DTD (Cimpoi et al., 2014): nearest-neighbor split (5 / 47)					
Original CLIP	59.4	42.3	25.5	79.2	54.3
SLUG	25.6	27.7	2.5	78.0	27.2
CLIPErase	1.3	19.7	12.0	82.0	21.9
DAMP	41.1	43.1	26.2	71.6	54.1
LIFT	1.1	44.0	28.7	42.8	54.2
DomainNet (Peng et al., 2019): real images (10 / 40)					
Original CLIP	96.0	95.6	95.5	96.3	54.3
SLUG	2.9	37.3	37.4	93.1	9.3
CLIPErase	3.5	7.6	6.6	57.6	4.1
DAMP	19.7	96.6	96.6	88.4	52.8
LIFT	7.5	96.3	96.3	62.3	53.4
COCO (Lin et al., 2014): object crops (5 / 80)					
Original CLIP	90.3	75.0	73.2	94.4	54.3
SLUG	0.9	58.9	46.6	93.9	29.3
CLIPErase	0.0	2.0	0.1	29.6	0.9
DAMP	16.4	75.2	74.1	84.6	53.6
LIFT	0.4	75.2	74.3	55.7	52.7

RESISC45. The retain-inactive adapter matters on three of the four requests: standard LoRA lowers Near-Retain accuracy by 3.3–6.6 points on Food-101, RESISC45, and DomainNet sketch. Overall, replacing any single component costs either retention or resistance to recovery, and only the full LIFT design keeps both. Appendices E–I add a second backbone (CLIP ViT-L/14), domain-selective and sequential requests, an alternative anchor pool, and stronger recovery attacks.

7 Conclusion

We presented LIFT, a utility-preserving method for CLIP class unlearning. LIFT treats unlearning as controlled feature relocation: moving forget features toward cluster-specific distributions of non-forget anchor features. Across challenging fine-grained datasets, LIFT achieves strong target suppression, reduces few-shot recovery from frozen visual features, and importantly preserves OOD zero-shot accuracy. These results suggest that effective CLIP unlearning should not only move target features away from their original labels, but also control where those features are moved.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing and improve grammar. We also used them to refine the format of tables and figures. All research ideas, methods, experiments, analyses, and substantive writing were carried out by the authors without AI/LLM assistance.

ETHICS STATEMENT

This work studies class-level unlearning for CLIP models. The intended use of LIFT is to support responsible model maintenance by reducing a model's reliance on designated target classes while preserving the general zero-shot utility of the original model. Such methods may be useful when practitioners need to update pretrained vision-language models in response to data governance requirements, dataset curation decisions, or deployment-specific removal requests.

All experiments are conducted on standard public research benchmarks and pretrained CLIP checkpoints. We do not collect new data, use human subjects, or perform human annotation. Our evaluation focuses on model behavior, including target-class suppression, few-shot recoverability, and OOD zero-shot transfer.

Unlearning methods may support responsible removal of unwanted class behavior, but they should not be interpreted as formal guarantees that all information about a concept has been erased. We therefore present LIFT as an empirical unlearning method and evaluate it under multiple recovery protocols. All datasets and compared methods in the paper are used in accordance with their respective terms of use and licenses.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of our work. Section 3 describes the method, and Sections 5.1 and 6 give the detailed implementation settings. Section 5 and Appendix A report the ablations of the class-transport distance, the aggregation unit, the loss terms and the adapter, together with the complete tables. Appendices C and D detail the sensitivity to the loss weights, Appendix H the choice of the anchor pool. We commit to providing the source code of our implementation upon publication to facilitate reproduction of our results.

REFERENCES

Jacopo Bonato, Marco Cotogni, and Luigi Sabetta. Is retain set all you need in machine unlearning? restoring performance of unlearned models with out-of-distribution images. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2024.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021.

Zikui Cai, Yaoteng Tan, and M Salman Asif. Targeted unlearning with single layer unlearning gradient. *ICML*, 2025.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7766–7775, 2023.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Rishav Chourasia and Neil Shah. Forget unlearning: Towards true data-deletion in machine learning. In *International conference on machine learning*, pp. 6028–6073. PMLR, 2023.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7210–7217, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Jimmy Z Di, Jack Douglas, Jayadev Acharya, Gautam Kamath, and Ayush Sekhari. Hidden poison: Machine unlearning enables camouflaged poisoning attacks. In *NeurIPS ML Safety Workshop*, 2022.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*, volume 2024, pp. 53643–53673, 2024.

Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards llm unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. *arXiv preprint arXiv:2502.05374*, 2025.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12043–12051, 2024.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

SeungBum Ha, Saerom Park, and Sung Whan Yoon. Unlearning's blind spots: Over-unlearning and prototypical relearning attack. *arXiv preprint arXiv:2506.01318*, 2025.

Arman Hatami, Romina Aalishah, and Ilya E Monosov. Class unlearning via depth-aware removal of forget-specific directions. *arXiv preprint arXiv:2604.15166*, 2026.

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 497–519. IEEE, 2025.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. In *International Conference on Learning Representations*, volume 2025, pp. 8857–8888, 2025.

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)*, volume 2, 2011.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Qing Li, Jiahui Geng, Derui Zhu, Fengyu Cai, Chenyang Lyu, and Fakhri Karray. Sauce: Selective concept unlearning in vision-language models with sparse autoencoders. *arXiv preprint arXiv:2503.14530*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

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*, pp. 722–729. IEEE, 2008.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415. IEEE, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Weiqi Wang, Zhiyi Tian, Chenhan Zhang, and Shui Yu. Machine unlearning: A comprehensive survey. *arXiv preprint arXiv:2405.07406*, 2024a.

Yihan Wang, Yiwei Lu, Guojun Zhang, Franziska Boenisch, Adam Dziedzic, Yaoliang Yu, and Xiao-Shan Gao. Muc: Machine unlearning for contrastive learning with black-box evaluation. *arXiv preprint arXiv:2406.03603*, 2024b.

Tianyu Yang, Lisen Dai, Xiangqi Wang, Minhao Cheng, Yapeng Tian, and Xiangliang Zhang. Cliperase: Efficient unlearning of visual-textual associations in clip. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30438–30452, 2025.

Zeliang Zhang, Gaowen Liu, Charles Fleming, Ramana Rao Kompella, and Chenliang Xu. Targeted forgetting of image subgroups in clip models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9870–9880, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022.

Yu Zhou, Dian Zheng, Qijie Mo, Renjie Lu, Kun-Yu Lin, and Wei-Shi Zheng. Decoupled distillation to erase: A general unlearning method for any class-centric tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20350–20359, 2025.

APPENDIX

A WHOLE-GROUP FORGETTING

This appendix contains the per-dataset and per-shot versions of Table 1, the aggregation, loss-term and adapter ablations summarised in Section 5, and the robustness checks of that section: recovery attacks and the sensitivity to the cluster and anchor counts.

Table 5: OOD zero-shot accuracy after unlearning (%, ↑, 3-seed mean). Original CLIP (Radford et al., 2021) is the non-unlearned reference. Baselines are DAMP (Hatami et al., 2026), CLIPErase (Yang et al., 2025), and SLUG (Cai et al., 2025). The OOD datasets are excluded from every forget pool, retain-distillation pool, and anchor pool. Best unlearned result in each column is in **bold**.

Method	Forget target: Cars				Forget target: CUB				Forget target: Dogs			
	Flowers	Aircraft	UCF101	Avg.	Flowers	Aircraft	UCF101	Avg.	Flowers	Aircraft	UCF101	Avg.
Original CLIP (Radford et al., 2021)	70.73	24.81	67.46	54.33	70.73	24.81	67.46	54.33	70.73	24.81	67.46	54.33
SLUG (Cai et al., 2025)	34.02	16.37	61.93	37.44	2.88	7.38	55.85	22.04	1.95	2.78	10.34	5.02
CLIPErase (Yang et al., 2025)	8.50	1.01	14.06	7.86	1.87	1.23	18.08	7.06	1.64	1.01	4.34	2.33
DAMP (Hatami et al., 2026)	34.32	10.26	60.66	35.08	26.08	11.17	56.79	31.35	50.41	15.84	61.30	42.52
LIFT (ours)	**70.39**	**24.32**	**67.47**	**54.06**	**68.40**	**24.70**	**67.37**	**53.49**	**70.51**	**24.80**	**66.95**	**54.09**

Table 6: Few-shot Gaussian Discriminant Analysis (GDA) recovery accuracy on forget classes (%, ↓, 3-seed mean). Column headers indicate the number of support shots. Original CLIP (Radford et al., 2021) is the non-unlearned reference. Baselines are SLUG (Cai et al., 2025), CLIPErase (Yang et al., 2025), and DAMP (Hatami et al., 2026). Best unlearned result per column is in **bold**.

Method	1	2	4	8	16	Avg.
Forget target: Cars						
Original CLIP (Radford et al., 2021)	32.65	51.62	64.50	72.42	77.28	59.69
SLUG (Cai et al., 2025)	30.39	47.69	60.66	69.13	74.32	56.44
CLIPErase (Yang et al., 2025)	4.60	7.16	11.74	16.86	21.85	12.44
DAMP (Hatami et al., 2026)	10.13	18.12	28.20	37.73	45.25	27.89
LIFT (ours)	**2.05**	**3.16**	**4.72**	**6.15**	**8.75**	**4.97**
Forget target: CUB						
Original CLIP (Radford et al., 2021)	33.67	51.19	62.43	69.42	74.25	58.19
SLUG (Cai et al., 2025)	31.33	46.62	58.94	67.13	72.26	55.26
CLIPErase (Yang et al., 2025)	4.66	7.44	11.34	16.46	21.03	12.19
DAMP (Hatami et al., 2026)	15.98	25.55	34.18	42.29	47.75	33.15
LIFT (ours)	**3.02**	**5.08**	**8.20**	**11.61**	**16.81**	**8.94**
Forget target: Dogs						
Original CLIP (Radford et al., 2021)	16.58	32.57	44.00	53.19	60.78	41.42
SLUG (Cai et al., 2025)	10.80	18.03	27.09	36.41	43.68	27.20
CLIPErase (Yang et al., 2025)	5.65	8.20	12.19	17.07	21.89	13.00
DAMP (Hatami et al., 2026)	6.97	11.85	16.51	21.84	25.56	16.55
LIFT (ours)	**3.35**	**5.51**	**8.15**	**11.97**	**17.05**	**9.21**

Aggregation unit in class transport. Table 7 studies whether class transport should be applied to individual images or to groups of forget features. In the per-image variant, each forget image retrieves its own nearest retain anchors and is matched independently. This gives similar target zero-shot accuracy, but it leaves much more recoverable information in the visual representation. This is undesirable for unlearning: even if the zero-shot classifier no longer predicts the target class, a four-shot Gaussian Discriminant Analysis classifier can still recover the forgotten classes from the frozen features. Averaged over Dogs, CUB, and Cars, per-image matching gives 14.95% four-shot recovery, while the class-cluster design used by LIFT reduces it to 7.02%. This indicates that moving each image toward its own nearest anchors can lower the zero-shot score, but does not sufficiently weaken the shared class structure of the forget set. In contrast, LIFT groups forget images by class and visual cluster, then matches each group to a shared non-forget anchor distribution. This group-

level transport gives lower recovery and higher OOD zero-shot accuracy on average, supporting the class-cluster aggregation used in the final objective.

Table 7: Ablating the aggregation unit in class transport. *Per-image* matches each forget image to its nearest retain anchors. *Class-cluster* is the LIFT default: forget features are grouped by class and visual cluster, then each group is matched to a shared anchor distribution.

Metric	Aggregation	Cars	CUB	Dogs	Avg.
Target↓	Per-image	0.53	**0.51**	**0.74**	0.59
	Class-cluster	**0.46**	0.53	0.76	**0.58**
GDA@4↓	Per-image	11.78	15.62	17.44	14.95
	Class-cluster	**4.72**	**8.20**	**8.15**	**7.02**
OOD↑	Per-image	**54.12**	53.02	53.57	53.57
	Class-cluster	54.06	**53.49**	**54.09**	**53.88**

Loss-term ablation. Table 8 separates the roles of the three losses. Retain distillation alone keeps OOD zero-shot accuracy close to the original model, but it does not unlearn the target classes: target zero-shot accuracy remains high, and the forget classes are still easily recovered. Class transport alone changes the behavior sharply, reducing the average target accuracy from 57.15% to 0.86% and the average four-shot recovery from 53.90% to 5.17%. The retain loss plays a different role: when added to class transport, it improves average OOD zero-shot accuracy from 53.58% to 53.93%, showing that it helps preserve general CLIP utility. Finally, adding the forget-text regularizer gives the strongest target suppression, reducing average target accuracy further to 0.58%. Its recovery value is slightly higher than the two-term objective, indicating that this term mainly suppresses direct image-text similarity to the target class rather than serving as the main mechanism for reducing recoverable visual structure.

Adapter ablation. Table 8 compares the retain-inactive LoRA adapter used in LIFT with a standard LoRA adapter (Hu et al., 2022), while keeping the loss objective and all hyperparameters fixed. Both variants achieve very low target zero-shot accuracy, but LIFT is consistently lower: Target-ZS decreases from 0.79 to 0.76 on Dogs, from 0.55 to 0.53 on CUB, and from 0.53 to 0.46 on Cars. The gain is clearer for utility preservation: retain-inactive LoRA gives higher OOD-ZS across all three forget targets, with the largest improvement on CUB from 51.84 to 53.49.

Table 8: Top: loss-term ablation on the three forget targets; T = class transport, R = retain distillation, F = forget-text regularizer; Target-ZS = target zero-shot accuracy (↓), Recovery = four-shot GDA recovery (↓), OOD-ZS = mean OOD zero-shot accuracy (↑), Avg. over the targets. Bottom: adapter ablation, standard LoRA (Hu et al., 2022) versus the retain-inactive adapter of LIFT, same objective and hyperparameters.

	Target-ZS ↓				Recovery ↓				OOD-ZS ↑			
Objective	Cars	CUB	Dogs	Avg.	Cars	CUB	Dogs	Avg.	Cars	CUB	Dogs	Avg.
R	64.06	50.24	57.14	57.15	61.30	58.68	41.71	53.90	54.37	54.27	54.33	54.32
T	0.73	0.90	0.96	0.86	4.15	6.22	5.13	5.17	54.09	52.61	54.03	53.58
T+R	0.93	0.98	1.30	1.07	4.41	7.70	5.29	5.80	54.09	53.57	54.13	53.93
T+R+F (LIFT)	0.46	0.53	0.76	0.58	4.72	8.20	8.15	7.02	54.06	53.49	54.09	53.88

	Target-ZS ↓			OOD-ZS ↑		
Adapter	Cars	CUB	Dogs	Cars	CUB	Dogs
Standard LoRA (Hu et al., 2022)	0.53	0.55	0.79	53.09	51.84	53.38
Retain-inactive (LIFT)	**0.46**	**0.53**	**0.76**	**54.06**	**53.49**	**54.09**

Broader recovery attacks. Table 9 tests four recovery attacks: prompt changes, full-support linear probes, retrieval, and LoRA relearning. LIFT is less recoverable than DAMP under every attack and needs more relearning steps. CLIPErase is sometimes lower still, but its OOD accuracy is only

2.33–7.86% before any attack, so its general representation has already collapsed. SLUG shows the same pattern in a weaker form: its low Dogs numbers come with a pre-attack OOD accuracy of 5.02%, while on CUB and Cars its retrieval mAP (0.296 / 0.283) stays close to Original CLIP (0.328 / 0.311) and relearning reaches 25% within 17 and 4 steps.

Cluster and anchor choices. Table 10 shows stable target suppression and OOD utility across $K \in \{4, 8, 16, 32, 64\}$ and $M \in \{64, 128, 256, 512\}$. More anchors reduce recovery on CUB and Cars; $M = 256$ retains a smaller anchor budget with low recovery. OOD accuracy varies by at most 0.30 points across the tested cluster counts.

Table 9: Additional whole-group attacks. Prompt-ZS is mean accuracy ± standard deviation across up to 21 individual templates. Linear probe attack uses the full forget-training split. Retrieval is same-class image-to-image with 16 support queries per class. Steps counts how many LoRA updates needed to reach 25% recovery with four supported images per class. OOD is measured before the attack.

Method	Prompt-ZS↓	Linear↓	mAP↓	Steps↑	OOD↑
Dogs					
Original CLIP (Radford et al., 2021)	58.75 ± 1.72	63.08	0.133	0	54.33
SLUG (Cai et al., 2025)	1.28 ± 0.44	2.50	0.079	38	5.02
CLIPErase (Yang et al., 2025)	0.82 ± 0.05	1.23	0.034	131	2.33
DAMP (Hatami et al., 2026)	4.16 ± 0.50	29.89	0.049	13	42.52
LIFT	0.84 ± 0.09	13.88	0.024	55	54.09
CUB					
Original CLIP (Radford et al., 2021)	54.45 ± 1.20	62.07	0.328	0	54.33
SLUG (Cai et al., 2025)	3.66 ± 0.86	13.25	0.296	17	22.04
CLIPErase (Yang et al., 2025)	0.53 ± 0.01	1.13	0.032	315	7.06
DAMP (Hatami et al., 2026)	6.20 ± 1.14	27.36	0.132	11	31.35
LIFT	0.42 ± 0.09	4.30	0.024	52	53.49
Cars					
Original CLIP (Radford et al., 2021)	65.03 ± 1.04	74.64	0.311	0	54.33
SLUG (Cai et al., 2025)	11.50 ± 3.39	61.08	0.283	4	37.44
CLIPErase (Yang et al., 2025)	0.63 ± 0.03	0.85	0.030	> 1000	7.86
DAMP (Hatami et al., 2026)	6.59 ± 0.78	36.19	0.077	10	35.08
LIFT	0.47 ± 0.06	2.96	0.015	49	54.06

Table 10: Sensitivity analysis on cluster and anchors choices. Each cell is Target-ZS / GDA@4 / OOD-ZS (%, three-seed means, lower is better for the first two values, higher for OOD.)

Value	Dogs	CUB	Cars
Clusters K; anchors $M = 256$			
4	0.66 / 8.29 / 54.07	0.51 / 7.69 / 53.54	0.43 / 3.99 / 54.16
8	0.57 / 8.32 / 54.14	0.45 / 8.45 / 53.52	0.44 / 4.15 / 54.24
16	0.76 / 8.15 / 54.09	0.53 / 8.20 / 53.49	0.46 / 4.72 / 54.06
32	0.85 / 8.19 / 54.20	0.50 / 8.63 / 53.48	0.51 / 4.90 / 54.10
64	1.00 / 9.21 / 54.13	0.61 / 9.72 / 53.67	0.50 / 5.08 / 54.18
Anchors M; clusters $K = 16$			
64	0.81 / 8.45 / 54.07	0.51 / 10.51 / 53.47	0.61 / 6.23 / 54.05
128	0.77 / 8.42 / 54.11	0.52 / 9.72 / 53.43	0.55 / 5.36 / 54.07
256	0.76 / 8.15 / 54.09	0.53 / 8.20 / 53.49	0.46 / 4.72 / 54.06
512	0.78 / 8.26 / 54.19	0.48 / 6.16 / 53.49	0.31 / 4.31 / 54.32

B SELECTIVE FORGETTING

Tables 11 and 12 are the complete versions of Table 4, with standard deviations over three seeds, 16-shot recovery and COCO-5k text-to-image retrieval (R@1).

Table 11: Selective forgetting of every Stanford Cars manufacturer with at least ten models. These classes (forget vs retain) are highly correlated in semantic understanding. DAMP and SLUG use all remaining classes of the dataset as retain data, as in their papers; CLIPErase uses the retain set of LIFT.

Method	Forget↓	Retain↑	Near↑	GDA@4↓	GDA@16↓	OOD↑	COCO↑
Audi (14 / 196)							
Original CLIP (Radford et al., 2021)	55.2	66.2	74.5	48.1	61.3	54.3	33.4
SLUG (Cai et al., 2025)	1.5 ± 0.0	45.5 ± 0.1	40.8 ± 0.2	47.4 ± 0.0	60.4 ± 0.0	51.9 ± 0.0	28.0 ± 0.0
CLIPErase (Yang et al., 2025)	0.0 ± 0.0	12.0 ± 6.7	7.9 ± 5.5	49.2 ± 0.9	60.6 ± 1.2	13.9 ± 4.3	17.3 ± 3.8
DAMP (Hatami et al., 2026)	54.1 ± 0.0	65.8 ± 0.0	74.4 ± 0.0	47.2 ± 0.0	61.7 ± 0.0	54.4 ± 0.0	33.7 ± 0.0
LIFT	0.3 ± 0.0	65.9 ± 0.0	73.1 ± 0.1	15.7 ± 0.5	24.9 ± 0.4	54.1 ± 0.1	33.2 ± 0.0
BMW (13 / 196)							
Original CLIP (Radford et al., 2021)	62.6	65.6	66.9	51.3	68.0	54.3	33.4
SLUG (Cai et al., 2025)	0.3 ± 0.1	45.7 ± 0.1	43.5 ± 0.1	50.6 ± 0.0	66.1 ± 0.1	51.8 ± 0.0	26.0 ± 0.0
CLIPErase (Yang et al., 2025)	0.0 ± 0.0	6.6 ± 0.8	2.9 ± 3.9	55.4 ± 1.8	67.9 ± 2.0	11.9 ± 3.5	13.6 ± 3.0
DAMP (Hatami et al., 2026)	57.2 ± 0.0	65.6 ± 0.0	67.1 ± 0.0	50.3 ± 0.0	67.3 ± 0.0	53.9 ± 0.0	33.2 ± 0.0
LIFT	0.0 ± 0.0	65.3 ± 0.1	66.8 ± 0.2	13.2 ± 0.5	20.8 ± 0.4	54.2 ± 0.1	33.2 ± 0.1
Chevrolet (22 / 196)							
Original CLIP (Radford et al., 2021)	47.2	67.7	58.9	61.0	71.4	54.3	33.4
SLUG (Cai et al., 2025)	1.0 ± 0.0	44.2 ± 0.2	30.3 ± 0.2	60.2 ± 0.0	70.0 ± 0.1	51.1 ± 0.2	23.6 ± 0.1
CLIPErase (Yang et al., 2025)	0.0 ± 0.0	3.4 ± 0.4	2.0 ± 1.3	57.9 ± 0.2	71.3 ± 0.3	9.7 ± 1.7	11.8 ± 1.4
DAMP (Hatami et al., 2026)	43.2 ± 0.1	67.6 ± 0.0	59.4 ± 0.0	60.2 ± 0.0	70.1 ± 0.0	53.6 ± 0.0	33.5 ± 0.0
LIFT	1.2 ± 0.1	67.5 ± 0.1	60.5 ± 0.5	16.9 ± 0.2	24.9 ± 0.6	54.1 ± 0.0	32.9 ± 0.0
Dodge (15 / 196)							
Original CLIP (Radford et al., 2021)	56.6	66.1	62.1	71.4	84.5	54.3	33.4
SLUG (Cai et al., 2025)	1.2 ± 0.1	44.2 ± 0.1	36.9 ± 0.1	70.4 ± 0.1	83.2 ± 0.1	52.1 ± 0.0	27.4 ± 0.1
CLIPErase (Yang et al., 2025)	0.0 ± 0.0	6.4 ± 2.2	1.2 ± 1.6	66.4 ± 1.0	79.0 ± 0.9	15.8 ± 1.7	15.0 ± 0.9
DAMP (Hatami et al., 2026)	52.8 ± 0.1	65.7 ± 0.0	61.7 ± 0.0	69.1 ± 0.0	83.1 ± 0.0	53.8 ± 0.0	33.3 ± 0.0
LIFT	0.3 ± 0.0	65.9 ± 0.1	59.9 ± 0.3	18.0 ± 0.9	29.2 ± 0.9	54.3 ± 0.1	32.8 ± 0.0
Ford (12 / 196)							
Original CLIP (Radford et al., 2021)	69.7	65.1	44.4	77.2	88.6	54.3	33.4
SLUG (Cai et al., 2025)	0.9 ± 0.1	42.5 ± 0.1	17.2 ± 0.2	76.4 ± 0.0	86.9 ± 0.0	51.3 ± 0.0	25.0 ± 0.0
CLIPErase (Yang et al., 2025)	0.1 ± 0.1	12.1 ± 6.5	4.3 ± 3.3	80.4 ± 1.2	89.1 ± 0.5	25.8 ± 18.5	19.6 ± 7.6
DAMP (Hatami et al., 2026)	65.7 ± 0.1	65.3 ± 0.0	44.8 ± 0.1	76.8 ± 0.0	88.2 ± 0.0	54.1 ± 0.0	33.7 ± 0.0
LIFT	0.2 ± 0.0	65.0 ± 0.1	45.4 ± 0.5	18.7 ± 0.6	29.0 ± 0.7	54.1 ± 0.1	33.1 ± 0.0
Hyundai (11 / 196)							
Original CLIP (Radford et al., 2021)	45.3	66.6	69.4	75.1	86.0	54.3	33.4
SLUG (Cai et al., 2025)	2.9 ± 0.0	43.6 ± 0.0	28.2 ± 0.1	75.1 ± 0.0	84.6 ± 0.0	53.0 ± 0.0	28.6 ± 0.0
CLIPErase (Yang et al., 2025)	0.0 ± 0.0	12.4 ± 0.9	3.3 ± 3.9	70.6 ± 1.9	82.0 ± 1.0	15.8 ± 4.7	18.3 ± 3.0
DAMP (Hatami et al., 2026)	40.3 ± 0.1	66.9 ± 0.0	71.0 ± 0.1	73.4 ± 0.0	85.0 ± 0.0	54.2 ± 0.0	33.5 ± 0.0
LIFT	0.0 ± 0.0	65.6 ± 0.0	64.3 ± 0.5	20.3 ± 0.2	30.5 ± 1.2	54.2 ± 0.1	32.6 ± 0.0

C SENSITIVITY TO THE LOSS WEIGHTS

Table 13 varies one loss weight at a time on whole-group forgetting. The retain weight λ_2 gives stable results: OOD-ZS rises slightly with it and barely improves beyond 3.0, while target-ZS and recovery move by at most 0.12 and 0.5 points over the whole range, so LIFT is not sensitive to this weight. The forget-text weight λ_1 is a monotone trade-off between forgetting and recoverability: raising it from 0.5 to 4.0 lowers the average target-ZS from 0.93% to 0.27% but raises four-shot recovery from 5.97% to 10.32%, because the extra forgetting pushes features into directions that a few-shot classifier can re-separate. The value 1.0 is the smallest weight that keeps target-ZS below 1% on all three families while average recovery stays within 1.5 points of its minimum. No other setting improves all three metrics at once. These runs use the retain configuration of the selective experiments, so their reference row differs slightly from Table 1 (by at most 0.12 points of target-ZS, 0.31 of OOD-ZS, and 0.62 of GDA@4).

Table 12: Selective forgetting across visual domains. Retain data as in Table 11.

Method	Forget↓	Retain↑	Near↑	GDA@4↓	GDA@16↓	OOD↑	COCO↑
Food-101 (Bossard et al., 2014): related dishes (10 / 101)							
Original CLIP (Radford et al., 2021)	80.6	86.5	85.9	95.5	97.7	54.3	33.4
SLUG (Cai et al., 2025)	2.7 ± 0.1	25.8 ± 0.3	20.8 ± 0.3	94.6 ± 0.0	97.4 ± 0.0	28.0 ± 0.1	8.2 ± 0.1
CLIPErase (Yang et al., 2025)	0.0 ± 0.0	1.1 ± 0.0	0.0 ± 0.0	24.6 ± 0.6	39.1 ± 0.5	0.8 ± 0.1	0.0 ± 0.0
DAMP (Hatami et al., 2026)	67.9 ± 0.1	86.8 ± 0.0	86.4 ± 0.0	94.3 ± 0.0	97.3 ± 0.0	54.1 ± 0.0	33.5 ± 0.0
LIFT	1.1 ± 0.0	86.3 ± 0.0	85.1 ± 0.1	29.4 ± 0.0	46.6 ± 0.7	54.5 ± 0.0	31.4 ± 0.1
RESISC45 (Cheng et al., 2017): related scenes (5 / 45)							
Original CLIP (Radford et al., 2021)	55.5	62.4	61.6	97.2	98.8	54.3	33.4
SLUG (Cai et al., 2025)	22.1 ± 0.0	55.2 ± 0.0	49.9 ± 0.0	96.9 ± 0.0	98.9 ± 0.0	45.5 ± 0.0	27.2 ± 0.0
CLIPErase (Yang et al., 2025)	6.7 ± 9.4	5.9 ± 2.6	0.0 ± 0.1	63.5 ± 5.5	80.1 ± 3.7	17.3 ± 9.3	19.4 ± 6.6
DAMP (Hatami et al., 2026)	28.1 ± 0.1	63.5 ± 0.0	62.9 ± 0.0	94.0 ± 0.0	98.1 ± 0.0	54.3 ± 0.0	33.6 ± 0.0
LIFT	0.0 ± 0.0	62.3 ± 0.1	58.8 ± 0.0	49.3 ± 0.4	68.3 ± 0.5	54.3 ± 0.0	32.6 ± 0.0
EuroSAT (Helber et al., 2019): land-cover classes (3 / 10)							
Original CLIP (Radford et al., 2021)	65.5	41.9	49.2	90.4	94.7	54.3	33.4
SLUG (Cai et al., 2025)	21.5 ± 0.1	34.1 ± 0.0	45.5 ± 0.0	90.4 ± 0.0	94.7 ± 0.0	49.1 ± 0.0	29.2 ± 0.0
CLIPErase (Yang et al., 2025)	0.0 ± 0.0	14.3 ± 0.0	0.0 ± 0.0	77.6 ± 5.2	88.7 ± 1.5	20.3 ± 7.9	9.6 ± 2.3
DAMP (Hatami et al., 2026)	22.1 ± 0.1	43.1 ± 0.0	50.5 ± 0.0	80.3 ± 0.0	86.9 ± 0.0	53.9 ± 0.0	33.1 ± 0.0
LIFT	0.4 ± 0.1	44.1 ± 0.3	51.9 ± 0.2	53.7 ± 0.5	64.7 ± 0.8	54.2 ± 0.0	32.8 ± 0.0
DTD (Cimpoi et al., 2014): nearest-neighbor split (5 / 47)							
Original CLIP (Radford et al., 2021)	59.4	42.3	25.5	79.2	93.1	54.3	33.4
SLUG (Cai et al., 2025)	25.6 ± 0.0	27.7 ± 0.0	2.5 ± 0.0	78.0 ± 0.0	92.3 ± 0.0	27.2 ± 0.0	9.3 ± 0.0
CLIPErase (Yang et al., 2025)	1.3 ± 0.5	19.7 ± 4.5	12.0 ± 1.8	82.0 ± 1.4	89.4 ± 1.8	21.9 ± 1.2	6.7 ± 1.6
DAMP (Hatami et al., 2026)	41.1 ± 0.0	43.1 ± 0.0	26.2 ± 0.0	71.6 ± 0.1	87.2 ± 0.0	54.1 ± 0.0	33.3 ± 0.0
LIFT	1.1 ± 0.0	44.0 ± 0.1	28.7 ± 0.3	42.8 ± 0.1	68.1 ± 0.7	54.2 ± 0.0	30.2 ± 0.0
DomainNet (Peng et al., 2019): real images (10 / 40)							
Original CLIP (Radford et al., 2021)	96.0	95.6	95.5	96.3	97.7	54.3	33.4
SLUG (Cai et al., 2025)	2.9 ± 0.2	37.3 ± 0.3	37.4 ± 0.4	93.1 ± 0.0	96.9 ± 0.0	9.3 ± 0.1	0.2 ± 0.0
CLIPErase (Yang et al., 2025)	3.5 ± 0.3	7.6 ± 2.6	6.6 ± 3.3	57.6 ± 1.6	64.7 ± 0.1	4.1 ± 2.2	0.2 ± 0.2
DAMP (Hatami et al., 2026)	19.7 ± 0.1	96.6 ± 0.0	96.6 ± 0.0	88.4 ± 0.0	95.4 ± 0.0	52.8 ± 0.0	33.0 ± 0.0
LIFT	7.5 ± 0.9	96.3 ± 0.0	96.3 ± 0.1	62.3 ± 0.4	79.9 ± 0.1	53.4 ± 0.1	28.9 ± 0.0
COCO (Lin et al., 2014): object crops (5 / 80)							
Original CLIP (Radford et al., 2021)	90.3	75.0	73.2	94.4	96.5	54.3	33.4
SLUG (Cai et al., 2025)	0.9 ± 0.1	58.9 ± 0.3	46.6 ± 0.5	93.9 ± 0.0	96.0 ± 0.0	29.3 ± 0.3	16.4 ± 0.2
CLIPErase (Yang et al., 2025)	0.0 ± 0.0	2.0 ± 0.9	0.1 ± 0.1	29.6 ± 3.6	35.4 ± 10.0	0.9 ± 0.2	0.0 ± 0.0
DAMP (Hatami et al., 2026)	16.4 ± 0.5	75.2 ± 0.0	74.1 ± 0.0	84.6 ± 0.0	93.2 ± 0.0	53.6 ± 0.0	33.0 ± 0.0
LIFT	0.4 ± 0.0	75.2 ± 0.0	74.3 ± 0.0	55.7 ± 0.3	72.8 ± 0.6	52.7 ± 0.1	30.0 ± 0.1

D Design ablations and retain-distillation weight under selective forgetting

The whole-group ablations of Section 5 cannot show what each component of LIFT does to the classes that must survive. Table 14 therefore repeats them on four selective requests with the Retain and Near-Retain columns. The class-cluster aggregation is what protects the retained classes: matching each forget image to its own anchors forgets equally well but costs 8.8–21.2 points of Near-Retain and up to 17.0 points of Retain. The MMD discrepancy is what keeps the forgotten classes hard to recover: pointwise cosine matching reaches the same target accuracy but recovery rises by 16–48 points, up to the level of Original CLIP on RESISC45, and pointwise MSE fails to forget on sketch (80.3%) and leaves Food-101 almost as recoverable as Original CLIP (89.9% four-shot recovery). The retain-inactive adapter matters when the retained classes share visual support with the forgotten ones: a standard LoRA costs 3.3–6.6 points of Near-Retain on Food-101, RESISC45 and sketch and is within 1.5 points on the small Audi request.

Table 15 repeats the λ_2 sweep on two selective requests, because the retain distillation could matter more when retained classes are close to the forgotten ones. It does not: across 0.5–4.0 every column moves by at most 1.5 points. On Food-101 the weight trades about one point of forgetting for about 1.5 points of Near-Retain and saturates at 3.0; on Audi it is flat. Retention on selective requests is set by the composition of the retain set and by the aggregation unit, not by this weight.

Table 13: Sensitivity of LIFT to the loss weights on whole-group forgetting (all classes of Dogs / CUB / Cars; mean over three seeds). One weight varies at a time; the row marked (used) is the setting of every experiment in the paper. Target-ZS and GDA@4 lower is better, OOD-ZS higher is better (Original CLIP (Radford et al., 2021): 54.33 on the same suite).

	Target-ZS ↓			OOD-ZS ↑			GDA@4 ↓		
Setting	Dogs	CUB	Cars	Dogs	CUB	Cars	Dogs	CUB	Cars
Retain-distillation weight λ_2 *(*$\lambda_1 = 1.0$*)*									
$\lambda_2 = 0.5$	0.77	0.55	0.47	54.12	53.35	54.04	8.44	8.48	4.65
$\lambda_2 = 1.0$	0.81	0.55	0.49	54.23	53.55	54.08	8.54	8.62	4.73
$\lambda_2 = 2.0$	0.85	0.56	0.51	54.30	53.70	54.07	8.67	8.80	4.79
$\lambda_2 = 3.0$ **(used)**	**0.88**	**0.56**	**0.53**	**54.32**	**53.80**	**54.11**	**8.65**	**8.82**	**4.88**
$\lambda_2 = 4.0$	0.89	0.53	0.53	54.34	53.85	54.11	8.75	8.98	4.93
Forget-text weight λ_1 *(*$\lambda_2 = 3.0$*)*									
$\lambda_1 = 0.5$	1.12	0.87	0.79	54.35	53.60	54.08	5.97	7.68	4.27
$\lambda_1 = 1.0$ **(used)**	**0.88**	**0.56**	**0.53**	**54.32**	**53.80**	**54.11**	**8.65**	**8.82**	**4.88**
$\lambda_1 = 2.0$	0.53	0.41	0.38	54.03	53.64	54.09	11.21	10.16	5.74
$\lambda_1 = 3.0$	0.37	0.32	0.34	53.97	53.55	54.10	12.00	10.53	6.28
$\lambda_1 = 4.0$	0.18	0.31	0.33	53.93	53.50	54.10	13.24	11.29	6.42

Table 14: Design ablations of LIFT on four selective requests (mean over three seeds; columns as in Table 4; Original CLIP (Radford et al., 2021) is the non-unlearned reference). Standard LoRA replaces the retain-inactive adapter; cosine and MSE replace the MMD discrepancy by pointwise matching to the selected anchors; per-image matches each forget image to its own anchors instead of each class-cluster group to an anchor distribution.

Variant	Forget	Retain	Near	GDA@4	OOD
Cars / Audi					
Original CLIP	55.2	66.2	74.5	48.1	54.3
LIFT (ours)	**0.3**	**65.9**	**73.1**	**15.7**	**54.1**
standard LoRA	0.2	65.8	73.8	16.1	54.2
cosine	0.2	65.3	71.3	34.7	54.1
MSE	0.6	66.0	75.3	37.7	54.2
per-image	0.2	63.2	61.7	24.8	53.7
Food-101					
Original CLIP	80.6	86.5	85.9	95.5	54.3
LIFT (ours)	**1.1**	**86.3**	**85.1**	**29.4**	**54.5**
standard LoRA	1.6	84.9	81.8	25.0	54.1
cosine	0.3	84.5	82.7	45.3	54.3
MSE	6.2	86.7	86.4	89.9	54.0
per-image	1.4	75.8	65.7	17.5	54.1

Variant	Forget	Retain	Near	GDA@4	OOD
RESISC45					
Original CLIP	55.5	62.4	61.6	97.2	54.3
LIFT (ours)	**0.0**	**62.3**	**58.8**	**49.3**	**54.3**
standard LoRA	0.2	61.7	54.6	46.8	53.6
cosine	0.0	60.3	53.6	97.5	54.4
MSE	0.1	62.8	60.4	96.6	54.5
per-image	0.0	58.0	50.0	65.5	54.1
DomainNet sketch					
Original CLIP	83.3	86.8	85.8	89.9	54.3
LIFT (ours)	**18.9**	**86.4**	**85.3**	**43.6**	**54.1**
standard LoRA	15.8	81.3	78.7	37.5	53.4
cosine	18.5	85.0	83.6	76.8	54.3
MSE	80.3	86.9	85.9	85.4	54.3
per-image	14.1	69.4	64.1	35.4	53.5

E SECOND BACKBONE

We conduct the whole-group experiments and four selective requests on OpenAI CLIP ViT-L/14 with every hyperparameter of LIFT unchanged from ViT-B/16 and every baseline at the settings of its paper; all methods use the same forget and retain splits as in the main experiments. Table 16 shows the our consistent improvement as on ViT-B/16 on every request compare to other unlearning methods. On whole-group forgetting LIFT reaches 0.3–0.9% target-ZS, keeps OOD-ZS within 0.7 points of the 62.4% of Original CLIP, and lowers four-shot recovery to 8–13%, whereas DAMP forgets partially and loses 7–14 points of OOD, SLUG keeps recovery at the level of Original CLIP and loses up to 45 points of OOD, and CLIPErase forgets by destroying the representation (OOD 4–11%). On the selective requests LIFT keeps Retain and Near-Retain at the level of Original CLIP with Forget below 1% on three of the four, and its weakest request, DomainNet sketch, is the same one as on ViT-B/16 (Forget 19.1%). The larger backbone consistently shows the ability of our method on preserving the ordinary CLIP utility.

Table 15: Retain-distillation weight λ_2 on two selective requests ($\lambda_1 = 1.0$; mean over three seeds; columns as in Table 14).

Setting	Forget	Retain	Near	GDA@4	OOD
Cars / Audi (14 of 196 models)					
$\lambda_2 = 0.5$	0.3	65.6	72.5	15.2	54.0
$\lambda_2 = 1.0$	0.3	65.8	72.8	15.3	54.1
$\lambda_2 = 2.0$	0.3	65.9	73.1	15.6	54.2
$\lambda_2 = 3.0$ **(used)**	**0.3**	**65.9**	**73.1**	**15.7**	**54.1**
$\lambda_2 = 4.0$	0.3	65.9	73.5	15.6	54.1
Food-101 (10 of 101 dishes)					
$\lambda_2 = 0.5$	0.5	85.7	83.8	29.2	54.1
$\lambda_2 = 1.0$	0.7	86.0	84.5	29.2	54.4
$\lambda_2 = 2.0$	0.9	86.2	84.9	29.3	54.5
$\lambda_2 = 3.0$ **(used)**	**1.1**	**86.3**	**85.1**	**29.4**	**54.5**
$\lambda_2 = 4.0$	1.4	86.4	85.3	29.1	54.5

Table 16: CLIP ViT-L/14 with unchanged hyperparameters (mean over three seeds). Top: whole-group forgetting of Dogs / CUB / Cars (Target-ZS ↓, GDA@4 ↓, OOD-ZS ↑; Original CLIP (Radford et al., 2021): target 68.1 / 59.7 / 76.7, GDA@4 59.2 / 72.3 / 75.0, OOD 62.4). Bottom: selective requests (columns as in Table 14; Original CLIP in each block header).

	Target-ZS ↓			GDA@4 ↓			OOD-ZS ↑		
Method	Dogs	CUB	Cars	Dogs	CUB	Cars	Dogs	CUB	Cars
SLUG (Cai et al., 2025)	1.4	1.5	7.0	52.3	69.5	72.8	17.4	29.4	42.1
CLIPErase (Yang et al., 2025)	0.8	0.5	0.6	9.6	19.8	9.8	4.0	6.8	10.5
DAMP (Hatami et al., 2026)	6.3	10.0	10.7	25.1	51.4	46.4	54.9	48.0	49.0
LIFT (ours)	**0.9**	**0.5**	**0.3**	**13.1**	**12.4**	**8.2**	**61.8**	**61.7**	**62.0**

Method	Forget	Retain	Near	GDA@4	OOD
Cars / Audi; Original CLIP 69.8 / 77.3 / 86.2 / 58.5 / 62.4					
SLUG (Cai et al., 2025)	2.9	57.5	51.2	57.8	58.6
CLIPErase (Yang et al., 2025)	0.0	40.9	18.0	53.2	50.8
DAMP (Hatami et al., 2026)	67.1	77.4	86.8	54.6	62.2
LIFT (ours)	**0.0**	**77.3**	**86.1**	**16.1**	**62.4**
Food-101; Original CLIP 88.2 / 91.2 / 89.9 / 97.5 / 62.4					
SLUG (Cai et al., 2025)	45.1	75.9	70.2	97.3	48.6
CLIPErase (Yang et al., 2025)	0.0	1.1	0.0	23.7	0.8
DAMP (Hatami et al., 2026)	75.5	91.6	90.7	96.5	62.3
LIFT (ours)	**0.6**	**91.4**	**90.0**	**26.7**	**62.1**
RESISC45; Original CLIP 77.3 / 65.9 / 57.1 / 99.0 / 62.4					
SLUG (Cai et al., 2025)	47.5	65.2	57.4	99.0	58.8
CLIPErase (Yang et al., 2025)	0.0	4.0	0.0	73.0	19.6
DAMP (Hatami et al., 2026)	48.5	66.5	58.3	96.1	62.3
LIFT (ours)	**0.4**	**68.3**	**60.6**	**45.2**	**62.2**
DomainNet sketch; Original CLIP 89.3 / 91.2 / 89.9 / 93.2 / 62.4					
SLUG (Cai et al., 2025)	73.0	90.4	89.6	93.1	49.2
CLIPErase (Yang et al., 2025)	1.5	52.3	46.5	50.4	39.8
DAMP (Hatami et al., 2026)	84.4	91.4	90.1	92.0	62.1
LIFT (ours)	**19.1**	**90.8**	**89.5**	**39.7**	**62.1**

F DOMAIN-SELECTIVE FORGETTING

A request may concern one visual domain of a class only. On DomainNet (Peng et al., 2019) we forget ten classes in the sketch domain while their photographs, paintings, and clipart stay in the retain data, and we evaluate the forgotten classes in every domain (Table 17). LIFT lowers the sketch accuracy of the forgotten classes from 83.3% to 18.9% while their photographs keep 94.1% of the original 96.0%, and the retained classes stay at the level of Original CLIP in every domain; paintings

and clipart, which resemble sketches, are partly affected (79.7% and 74.7%). DAMP barely forgets the sketches (68.2%); CLIPErase and SLUG forget them but also lose the photographs of the same classes (67.5% and 14.8%) and much of the retained sketches (42.8% and 12.9%), and their OOD accuracy falls to 26.3% and 4.4%.

Table 17: Forgetting ten DomainNet classes in the sketch domain only (mean over three seeds). Forget = zero-shot accuracy of the forgotten classes in each domain (↓ for sketch, ↑ for the other domains, which must be kept; the request is met only when the sketch column is low and the photo column stays high at the same time); Retain = accuracy of the retained classes in the sketch and photo domains (↑). Original CLIP (Radford et al., 2021): forget 83.3 / 96.0 / 85.7 / 89.5, retain 86.8 / 95.6, OOD 54.3.

	Forgotten classes				Retained classes		
Method	Sketch ↓	Photo ↑	Painting ↑	Clipart ↑	Sketch ↑	Photo ↑	OOD ↑
SLUG (Cai et al., 2025)	2.9	14.8	3.9	7.6	12.9	27.8	4.4
CLIPErase (Yang et al., 2025)	3.1	67.5	44.2	41.7	42.8	91.2	26.3
DAMP (Hatami et al., 2026)	68.2	96.0	84.5	85.6	87.0	95.2	54.2
LIFT (ours)	18.9	94.1	79.7	74.7	86.4	95.5	54.1

G Sequential selective requests

Table 18 applies three manufacturer requests to one model in sequence (Audi, then BMW, then Chevrolet), each stage starting from the previous edited model with the earlier targets excluded from the retain and anchor data, and compares the final model with a single joint request for the same 49 models. This mirrors the whole-group chain of Table 3 at the selective level. For LIFT the cumulative Forget stays below 0.5% at every stage, Retain and Near-Retain stay within 1.9 points of Original CLIP, and OOD accuracy does not move; the three sequential edits end within 1.4 points of the joint request on the retention columns, so the interference between correlated requests is bounded. DAMP never forgets the requested makes at any stage, and CLIPErase compounds its damage: after the first stage Retain is already 12.0%, and after three stages the retained models and OOD accuracy are largely gone. SLUG forgets the requested makes at every stage, but each edit removes a further part of the retained models: Retain falls from 45.5% after the first request to 15.6% after the third, Near-Retain from 40.8% to 9.2%, and OOD accuracy from 51.9% to 36.1%, while GDA@4 recovery stays at 47–52%; its joint request ends at the same point.

H Source of the anchor pool

Caltech-101 serves as the generic source of anchors and of the generic share of the retain distillation. Table 19 replaces it by Oxford Flowers, a much narrower dataset, on two requests, keeping everything else fixed; because Flowers is then no longer held out, OOD accuracy is measured on the two remaining OOD datasets for both pools. Every conclusion survives the swap: Forget moves by at most 0.2 points, Retain and Near-Retain by at most one point, and OOD accuracy by at most 0.7 points. Four-shot recovery is about two points higher with the narrower pool. The generic pool is therefore a convenience rather than a requirement; a more diverse pool buys about a point of retention and about two points of recovery resistance.

I Recovery attacks on selective requests

Table 20 repeats the attacks of Table 9 on three selective requests, with identical resources for every method. Under prompt changes, retrieval, and relearning with five hundred steps, LIFT keeps the forgotten classes far below Original CLIP while its OOD accuracy after the attack stays at the original level, and it needs 20–183 relearning steps to regain half of the original excess accuracy where DAMP starts from it. A fully supervised linear probe with hundreds of labelled images per forgotten class is the exception: it recovers most of the forgotten classes on Food-101 and DomainNet (95.6% and 96.7% against 98.3% and 99.1% for Original CLIP), and about two thirds of them on Audi

Table 18: Sequential selective requests on Stanford Cars (Audi, then BMW, then Chevrolet) and the joint request for all three makes (mean over three seeds). Forget is cumulative over the makes requested so far; columns as in Table 14. Original CLIP (Radford et al., 2021): stage 1 55.2 / 66.2 / 74.5 / 48.1 / 54.3; stage 2 58.7 / 66.4 / 78.4 / 50.1 / 54.3; stage 3 and joint 53.6 / 69.3 / 66.8 / 54.8 / 54.3.

Method	Forget	Retain	Near	GDA@4	OOD
Stage 1: Audi					
SLUG (Cai et al., 2025)	1.5	45.5	40.8	47.4	51.9
CLIPErase (Yang et al., 2025)	0.0	12.0	7.9	49.2	13.9
DAMP (Hatami et al., 2026)	54.1	65.8	74.4	47.2	54.4
LIFT (ours)	**0.3**	**65.9**	**73.1**	**15.7**	**54.1**
Stage 2: Audi + BMW					
SLUG (Cai et al., 2025)	0.6	34.3	29.0	48.8	47.2
CLIPErase (Yang et al., 2025)	0.0	12.0	6.4	48.2	9.6
DAMP (Hatami et al., 2026)	54.7	66.3	79.0	48.2	53.6
LIFT (ours)	**0.1**	**65.5**	**76.5**	**12.4**	**54.1**
Stage 3: Audi + BMW + Chevrolet					
SLUG (Cai et al., 2025)	1.1	15.6	9.2	52.4	36.1
CLIPErase (Yang et al., 2025)	0.4	8.0	5.6	27.5	7.5
DAMP (Hatami et al., 2026)	49.1	68.7	66.7	52.5	52.8
LIFT (ours)	**0.4**	**67.9**	**65.7**	**12.9**	**54.1**
Joint request: the same three makes at once					
SLUG (Cai et al., 2025)	1.5	16.1	11.0	51.7	36.0
CLIPErase (Yang et al., 2025)	0.0	0.9	0.2	27.3	2.6
DAMP (Hatami et al., 2026)	50.8	69.1	66.9	52.9	53.0
LIFT (ours)	**1.7**	**68.8**	**67.1**	**17.3**	**54.1**

Table 19: LIFT with Caltech-101 (used) or Oxford Flowers as the generic anchor and retain pool (mean over three seeds). OOD is the mean over FGVC Aircraft and UCF101 for both pools (Original CLIP (Radford et al., 2021): 46.1); other columns as in Table 14.

Generic pool	Forget	Retain	Near	GDA@4	OOD	COCO t2i R@1
Food-101 (10 of 101 dishes); Original CLIP 80.6 / 86.5 / 85.9 / 95.5 / 46.1 / 33.4						
Caltech-101 **(used)**	**1.1**	**86.3**	**85.1**	**29.4**	**46.1**	**31.4**
Oxford Flowers	1.3	85.3	84.5	31.1	45.4	30.9
Cars / Audi (14 of 196 models); Original CLIP 55.2 / 66.2 / 74.5 / 48.1 / 46.1 / 33.4						
Caltech-101 **(used)**	**0.3**	**65.9**	**73.1**	**15.7**	**46.0**	**33.2**
Oxford Flowers	0.3	65.5	73.7	17.9	46.2	33.3

(45.7% against 70.5%). LIFT therefore reduces recoverability under the evaluated attacks rather than erasing the classes, which is the claim made in the paper. DAMP is recovered completely by every attack, and the low numbers of CLIPErase come with a destroyed model: its OOD accuracy after relearning is 1–19%. SLUG behaves like DAMP under the linear probe and relearning (within two points of Original CLIP on all three requests), lowers retrieval mAP by 1–12 points, and lowers the prompt attack to 37–89%; its OOD accuracy after relearning falls to 34% on Food-101 and 19% on DomainNet.

Table 20: Recovery attacks on three selective requests (mean over three seeds; all in %). Linear = linear probe trained on the full forget training set; Prompt = best of 21 prompt templates; mAP = image retrieval of the forgotten classes; Relearn = accuracy after 500 steps of LoRA relearning with four images per class; OOD = OOD zero-shot accuracy after the relearning attack. Lower is better for the first four columns, higher for the last.

Method	Linear ↓	Prompt ↓	mAP ↓	Relearn ↓	OOD after ↑
Cars / Audi (14 of 196 models)					
Original CLIP (Radford et al., 2021)	70.5	57.9	27.4	61.6	52.0
SLUG (Cai et al., 2025)	71.3	36.8	26.5	59.7	51.4
CLIPErase (Yang et al., 2025)	67.4	10.8	23.2	55.5	12.5
DAMP (Hatami et al., 2026)	70.4	57.2	26.1	60.3	52.1
LIFT (ours)	**45.7**	**8.4**	**10.5**	**44.6**	**52.8**
Food-101 (10 of 101 dishes)					
Original CLIP (Radford et al., 2021)	98.3	97.9	81.5	98.0	54.0
SLUG (Cai et al., 2025)	98.2	89.0	76.5	97.8	34.2
CLIPErase (Yang et al., 2025)	9.1	10.4	14.6	46.8	1.0
DAMP (Hatami et al., 2026)	98.2	97.5	77.2	97.9	54.4
LIFT (ours)	**95.6**	**18.7**	**18.6**	**57.7**	**53.7**
DomainNet photos (10 of 40 classes)					
Original CLIP (Radford et al., 2021)	99.1	98.6	82.2	98.3	54.6
SLUG (Cai et al., 2025)	98.8	59.1	70.1	97.7	19.0
CLIPErase (Yang et al., 2025)	91.8	18.2	38.6	92.5	18.7
DAMP (Hatami et al., 2026)	98.5	96.2	72.1	97.9	54.4
LIFT (ours)	**96.7**	**17.2**	**40.5**	**76.9**	**53.8**