DIFFUNLEARN: SELECTIVELY UNLEARNABLE EXAMPLES VIA DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Images shared online can be used for model training without their owners' consent. To protect images from such unauthorized training, unlearnable examples add small perturbations before release to impair the generalization of models trained on the protected images. However, data owners may wish to protect only a target concept while retaining the training utility of non-target concepts within the same image. This motivates selective protection of concepts within the same image, a requirement that remains underexplored in existing approaches. To address this gap, we formulate this goal as selective unlearnability and propose DIFFUNLEARN, a diffusion-based framework for protecting a text-specified target concept. Our protection objective combines image-text alignment, target attention suppression, and visual preservation. To optimize this objective, we use a pretrained diffusion model to perform text-guided optimization of intermediate latents obtained through DDIM inversion of the input images. First, we alternate vision-language surrogate training with latent optimization, using image-text alignment during latent updates to encourage target-correlated shortcuts that reduce reliance on generalizable target features. Second, to provide spatial guidance, we suppress the diffusion model's target attention responses in regions associated with the target concept. Finally, the visual preservation loss constrains image distortion under a fixed perturbation budget. Experiments on multilabel classification show that DIFFUNLEARN suppresses learning of the selected target while largely preserving aggregate non-target training utility. We additionally design transfer evaluations across unseen models and downstream tasks using the same protected images without task-specific reoptimization. Code is available in `https://anonymous.4open.science/r/9F6B/`.

1 INTRODUCTION

Images shared online support visual representation learning at scale (Radford et al., 2021; Jia et al., 2021), but public visibility does not imply consent to model training. Unauthorized use can support classification, image-text matching, and generation that reproduces memorized training images (Carlini et al., 2023; Somepalli et al., 2023). Membership inference can also reveal whether an image was used for training (Duan et al., 2023). These concerns motivate protection that accompanies images when released.

Unlearnable examples (UEs) add protective perturbations before release, inducing models to fit misleading features that generalize poorly to clean images (Huang et al., 2021). Glaze and Anti-DreamBooth protect against unauthorized style mimicry and personalized synthesis (Shan et al., 2023; Van Le et al., 2023). Subsequent work improves robustness against adversarial training (Fu et al., 2022) and transferability (Ren et al., 2023), and broadens UE construction through convolutional perturbations (Sadasivan et al., 2023), label-agnostic clustering (Zhang et al., 2023), and multimodal contrastive learning (Liu et al., 2024b).

Protection has also become more configurable. One for All learns a universal generator for a target category across datasets and label systems (Chen et al., 2024), while T2UE generates perturbations from text descriptions (Ma et al., 2025). However, flexible target specification does not ensure that non-target concepts retain their training utility: perturbations can impair learning beyond the intended target.

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.

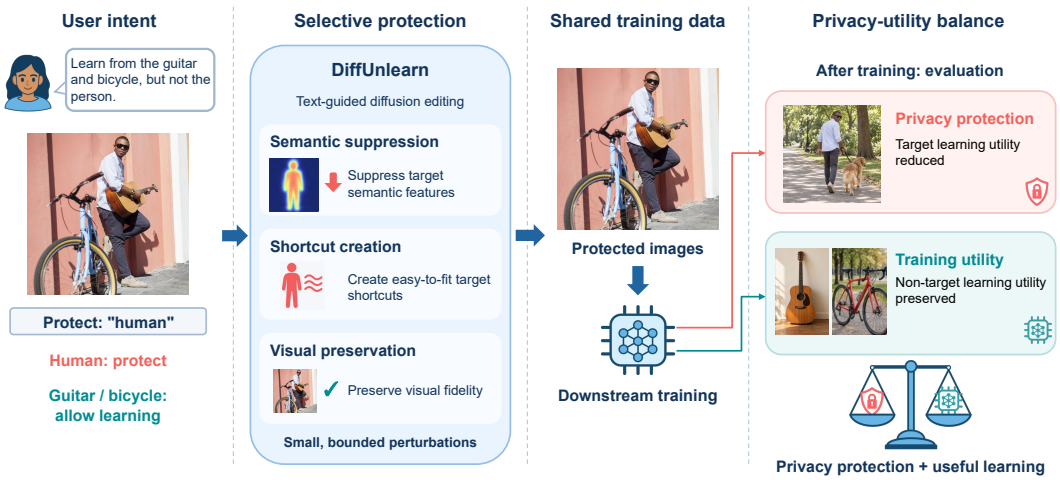

Figure 1: Selective unlearnability in images guided by user intent. DIFFUNLEARN aims to suppress learning of the target concept while preserving the training utility of non-target concepts and maintaining visual fidelity.

Data owners may instead wish to protect a particular concept while keeping others available for learning. In Figure 1, the owner of a photograph showing a person with a guitar and a bicycle may wish to protect the *human* concept while allowing the guitar and bicycle to support training. Impairing these other concepts would sacrifice utility the owner is willing to share. Can we suppress learning of a chosen concept while retaining the training utility of concepts that coexist with it?

We formulate this goal as *selective unlearnability in images*, using text to specify the *target concept*. Relative to training on the original images, training on protected images should reduce target performance on clean test images while largely preserving non-target performance. Image content remains visible; protection acts on subsequent learning from the released data.

Text-guided diffusion editing offers tools for translating the owner's choice into spatially guided protection: cross-attention connects prompt words to image structure (Hertz et al., 2023), and inversion initializes editing from an existing image (Song et al., 2021; Mokady et al., 2023). DiffAttack also generates transferable adversarial perturbations by optimizing intermediate diffusion latents (Chen et al., 2023). Building on these capabilities, DIFFUNLEARN uses Stable Diffusion (Rombach et al., 2022) and DDIM inversion to optimize an intermediate latent for selective protection.

Our guidance combines shortcut creation, target attention suppression, and visual preservation. We alternate vision-language pre-trained (VLP) surrogate training with latent optimization, using image-text alignment to encourage target-correlated shortcuts that generalize poorly to clean images. Target attention suppression penalizes diffusion attention in target-associated regions, while visual preservation limits changes to non-target regions and overall appearance under a fixed ℓ_∞ budget. Section 3 details the objectives and optimization.

We evaluate selective protection in image–text retrieval, transfer to classification, and behavior across unseen visual representations and training settings.

Our contributions are threefold:

- We formulate selective unlearnability in images as a form of proactive protection, explicitly coupling the suppression of a text-specified target concept with the preservation of non-target training utility within the same images.
- We propose DIFFUNLEARN, which adapts text-guided diffusion editing to the construction of unlearnable examples through shortcut creation, target attention suppression, and visual preservation under a bounded perturbation.
- We empirically demonstrate target suppression with largely preserved aggregate non-target utility in multilabel classification, and evaluate the transfer of target suppression to unseen training models and downstream tasks using the same protected images.

2 Related Work

Construction and configuration of unlearnable examples. UEs protect released images by inducing shortcuts that generalize poorly to clean data. Starting from alternating error-minimization with a surrogate (Huang et al., 2021), research has improved robustness against adversarial training (Fu et al., 2022), diversified construction through convolutional perturbations (Sadasivan et al., 2023) and label-agnostic clustering (Zhang et al., 2023), and extended protection to medical images (Sun et al., 2024) and multimodal learning (Liu et al., 2024b). PIL further improves efficiency with linear surrogates, attributing effectiveness to perturbation-induced linearization (Liu et al., 2026). Its partial-perturbation analysis varies the fraction of protected samples, rather than concepts within an image.

Beyond how perturbations are constructed, recent work considers what owners can configure. One for All provides reusable target-category generators across datasets and label systems (Chen et al., 2024); T2UE generates perturbations from text without original images (Ma et al., 2025); MMGuard controls multimodal fine-tuning protection through perturbation budgets, text triggers, and objectives (Zhao et al., 2026). These advances enable control over protection targets and strength, but leave open which coexisting concepts remain learnable. Our selective unlearnability formulation addresses this distinction by jointly requiring target suppression and non-target utility within the same images.

Diffusion-based image editing. Selective protection requires translating a textual target into image modifications while retaining surrounding content. Diffusion models generate images through iterative denoising (Ho et al., 2020), and latent diffusion enables text conditioning in a compressed image space (Rombach et al., 2022), and inversion extends editing to existing images: DDIM supports approximate inversion (Song et al., 2021), while Null-text inversion improves reconstruction fidelity (Mokady et al., 2023). Building on inversion, Plug-and-Play Diffusion Features preserves layout through source-feature injection (Tumanyan et al., 2023), and MasaCtrl maintains appearance through mutual self-attention control (Cao et al., 2023). DiffAttack further shows that inverted latents can generate transferable adversarial perturbations (Chen et al., 2023).

To connect latent editing to a chosen concept, spatial semantic guidance is also needed. Prompt-to-Prompt controls layout through word-specific cross-attention (Hertz et al., 2023); DAAM derives word–pixel attribution maps (Tang et al., 2023); further analysis distinguishes cross-attention object attribution from self-attention structure (Liu et al., 2024a). Together, inversion and target-token attention provide the editing variable and spatial reference for DiffUnlearn. We adapt these capabilities to impair subsequent target learning while preserving non-target utility, rather than inducing prediction errors on perturbed test inputs.

Transferability of unlearnable examples. Owners generally cannot anticipate which models or tasks will use their released images, so protection must transfer beyond its construction setting. For cross-model transfer, Unlearnable Clusters uses vision-language surrogates across models and label systems (Zhang et al., 2023), while PIL transfers linear-surrogate perturbations to deep classifiers (Liu et al., 2026). Multimodal Unlearnable Examples evaluates transfer between models (Liu et al., 2024b), and MMGuard improves it through surrogate ensembles (Zhao et al., 2026). One for All also broadens generator reuse across datasets and label systems (Chen et al., 2024).

Changing downstream tasks introduces a further challenge: protection must survive changes in learning objectives. TUE supports supervised and unsupervised training through contrastive learning and classwise separability (Ren et al., 2023). Multimodal Unlearnable Examples studies contrastive training with retrieval evaluation (Liu et al., 2024b), while MMGuard extends protection across vision-language fine-tuning tasks and recipes (Zhao et al., 2026). These studies motivate reusing protected images across learning settings. Our evaluation additionally asks whether both target suppression and non-target utility transfer, using the same images without task-specific reoptimization and matched clean-training baselines.

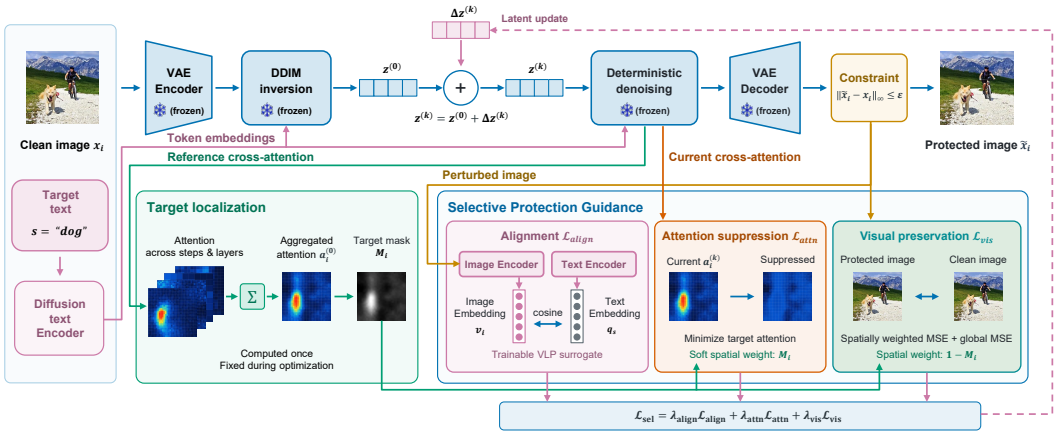

Figure 2: Overview of DIFFUNLEARN. Text-guided latent optimization combines image-text alignment, target attention suppression, and visual preservation to generate protected images under a bounded perturbation budget.

3 SELECTIVE UNLEARNABILITY IN IMAGES

3.1 PROBLEM FORMULATION

We consider protecting a text-specified concept while retaining the training utility of other concepts in the same images. Let $\mathcal{D} = \{x_i\}_{i=1}^N$ denote a training image set, where $x_i \in [0,1]^d$. The target text s specifies the target concept to be protected. Let $\mathcal{I}_s$ be the given index set of images containing the target concept. Given $(\mathcal{D}, s, \mathcal{I}_s)$, we produce a protected training set:

$$\widetilde{\mathcal{D}} = \{\widetilde{x}_i\}_{i=1}^N, \qquad \widetilde{x}_i = \begin{cases} x_i + \delta_i, & i \in \mathcal{I}_s, \\ x_i, & i \notin \mathcal{I}_s, \end{cases} \quad \|\delta_i\|_\infty \le \epsilon, \quad \widetilde{x}_i \in [0,1]^d. \tag{1}$$

Relative to training on $\mathcal{D}$ under the same setup, our aim is for training on $\widetilde{\mathcal{D}}$ to reduce target-concept performance on clean test images while largely preserving non-target performance, including for concepts that coexist with the target. The data owner generates the protected images without access to the model subsequently trained on the released data and cannot intervene in its subsequent training.

3.2 OUR METHOD

As shown in Figure 2, DIFFUNLEARN constructs protected images by editing an inverted diffusion latent. Target-token cross-attention provides spatial guidance for the attention suppression and visual preservation losses, while CLIP, used as our VLP surrogate, guides image-text alignment. Surrogate training alternates with latent optimization.

3.2.1 DIFFUSION INVERSION

We initialize latent optimization by mapping each image to an intermediate diffusion state. Let $\mathcal{E}$ and $\mathcal{G}$ be the VAE encoder and decoder, and let θ denote the diffusion-model parameters. For image x_i, we construct the diffusion text condition p_i from the target text s. We apply unconditional deterministic DDIM inversion to the encoded clean image and stop at an intermediate step τ (Song et al., 2021); p_i conditions the subsequent denoising. Denoting this mapping by Inv_θ^τ, the optimization is initialized and parameterized as:

$$z_{i,\tau}^{(0)} = \mathrm{Inv}_\theta^\tau(\mathcal{E}(x_i), \varnothing), \qquad z_{i,\tau}^{(k)} = z_{i,\tau}^{(0)} + \Delta z_{i,\tau}^{(k)}, \quad \Delta z_{i,\tau}^{(0)} = 0. \tag{2}$$

Here $\varnothing$ denotes empty-text conditioning during inversion. We optimize $\Delta z_{i,\tau}^{(k)}$ directly (Chen et al., 2023). This parameterization connects image modification to the target-token cross-attention along

the denoising trajectory, allowing both image-level and spatial attention objectives to guide the same latent variable.

At iteration k, deterministic denoising from $z_{i,\tau}^{(k)}$ under p_i followed by VAE decoding produces a decoded candidate image. Clipping this decoded candidate image to satisfy Equation 1 gives the protected candidate image $\widetilde{x}_i^{(k)}$ used in the image-based losses below.

3.2.2 Language-Guided Target Attention

We derive spatial guidance for the target concept from target attention responses during denoising. Let $\mathcal{K}_i$ denote the indices of the target tokens in p_i. During deterministic denoising initialized by $z_{i,\tau}^{(k)}$ under the diffusion text condition p_i, let $\mathcal{A}_\theta^\ell(z_{i,\tau}^{(k)}, p_i, t)$ denote the cross-attention response recorded at layer ℓ and diffusion step t. Its rows correspond to spatial latent positions and its columns correspond to prompt tokens. Let $\mathcal{S}$ and $\mathcal{L}$ denote the denoising steps and cross-attention layers. We aggregate only the target tokens $\mathcal{K}_i$ and upsample every response to the image resolution:

$$a_i^{(k)}(u) = \frac{1}{|\mathcal{S}|\,|\mathcal{L}|\,|\mathcal{K}_i|} \sum_{t\in\mathcal{S}} \sum_{\ell\in\mathcal{L}} \sum_{j\in\mathcal{K}_i} \mathrm{Up}\Big(\mathcal{A}_\theta^\ell(z_{i,\tau}^{(k)}, p_i, t)(:, j)\Big)\,(u). \tag{3}$$

The reference target attention response $a_i^{(0)}$ is obtained by denoising the unedited inverted latent $z_{i,\tau}^{(0)}$ and converted into a fixed soft target mask:

$$M_i(u) = \frac{a_i^{(0)}(u) - \min_v a_i^{(0)}(v)}{\max_v a_i^{(0)}(v) - \min_v a_i^{(0)}(v) + \zeta}, \tag{4}$$

where $\zeta > 0$ prevents division by zero. The target mask is computed once and detached during optimization, while the current target attention response $a_i^{(k)}$ remains differentiable with respect to the edited latent. Fixing M_i gives attention suppression and visual preservation a consistent spatial reference throughout optimization, even as the current target attention response changes. The mask provides soft weights without imposing a hard spatial constraint on the perturbation.

3.2.3 Selective Protection Guidance

Our selective protection guidance combines three complementary objectives. Image-text alignment guides the creation of target-related shortcuts, target attention suppression provides spatial guidance through the diffusion model, and visual preservation limits changes to image content. Together, these objectives guide latent optimization toward target protection while limiting interference with other content in the image.

Shortcut Creation. We guide perturbation generation using the shared target semantics. Image-text alignment is measured in the shared embedding space of the VLP surrogate. Let g_I and g_T denote the VLP encoders with normalized outputs, and let $\mathrm{sim}(u, v) = u^\top v$ denote cosine similarity. For the protected candidate image, write $v_i = g_I(\widetilde{x}_i)$ and $q_s = g_T(s)$, omitting the iteration index in the alignment formulation below. We align the protected image with the target text s by minimizing their cosine distance in the VLP embedding space:

$$\mathcal{L}_{\mathrm{align}} = 1 - \mathrm{sim}(v_i, q_s). \tag{5}$$

The target text provides a shared semantic direction across $\mathcal{I}_s$. With the VLP encoders fixed during latent updates, the alignment loss guides changes to the image through the latent perturbation. This objective introduces target-correlated cues that models can exploit when trained on the protected images. Such shortcuts tie the learned target association to perturbation cues that are absent from clean test images.

Target Attention Suppression. Image-text alignment provides a global semantic objective but does not specify where the modification should act. Target attention suppression weakens the diffusion model's cross-attention association between the target text and the regions identified by the initial target attention response. We implement this guidance by penalizing the current target attention response with the fixed spatial weights M_i:

$$\mathcal{L}_{\mathrm{attn}} = \frac{\sum_u M_i(u) a_i^{(k)}(u)}{\sum_u M_i(u) + \zeta}. \tag{6}$$

Alignment thus encourages compatibility with the target text in the VLP embedding space, while attention suppression discourages the diffusion model's target attention responses in the reference regions.

Visual Preservation. With $\delta_i = \widetilde{x}_i - x_i$, we penalize changes in low-target-attention regions and across the whole image:

$$\mathcal{L}_{\text{vis}} = \frac{\|(1 - M_i) \odot \delta_i\|_2^2}{\|1 - M_i\|_1 + \zeta} + \frac{1}{d}\|\delta_i\|_2^2. \tag{7}$$

The first term makes changes more costly at locations with weak initial target attention, supporting preservation of surrounding content. The second term penalizes distortion across the whole image, including high-attention regions where the first term assigns little weight. These soft penalties control distortion within the hard ℓ_∞ budget.

Combining image-text alignment, target attention suppression, and visual preservation, we obtain the selective protection objective:

$$\mathcal{L}_{\text{sel}} = \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}} + \lambda_{\text{vis}}\mathcal{L}_{\text{vis}}, \tag{8}$$

where the nonnegative weights control the relative contributions of the three objectives.

3.2.4 Alternating Optimization in Latent Space

Following the alternating optimization framework of unlearnable examples (Huang et al., 2021; Liu et al., 2024b), we alternate between surrogate training and latent perturbation updates. Updating the surrogate allows subsequent latent updates to account for its adaptation to the current protected images. During surrogate training, the current image perturbations are held fixed, and both VLP encoders are updated by minimizing the cosine alignment loss in Equation 5 on the current protected images and target text. During latent optimization, the VLP parameters are held fixed, and $\mathcal{L}_{\text{sel}}^{(k)}$ guides the perturbation updates. The diffusion model, VAE, and target mask remain fixed throughout both stages.

At each latent update, alignment and visual preservation losses are computed on $\widetilde{x}_i^{(k)}$, while attention suppression uses the current denoising trajectory. We backpropagate their combined loss through the frozen models to the latent perturbation and use an Adam-based optimizer to update $\Delta z_{i,\tau}$ directly. The clipping operation uses a straight-through gradient estimator.

Once optimization is complete, the final protected candidate images form the protected images in $\widetilde{\mathcal{D}}$ for subsequent training.

4 Experiments

We evaluate whether DiffUnlearn suppresses learning of the selected target while preserving non-target utility in image-text retrieval, whether these properties transfer to classification and unseen visual representations using the same protected images, and how the perturbations affect visual quality. The appendix provides implementation details (Appendix A.1), basic unlearnability results on CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) (Appendix A.2), the component-ablation protocol (Appendix A.3), and mathematical analysis (Appendix B).

4.1 Experimental Setup

Datasets and metrics. We use Flickr30k Entities and MS COCO (Plummer et al., 2015; Lin et al., 2014) for selective retrieval and cross-task classification. Models are trained on released images and evaluated on clean test images. We report classification accuracy, average precision (AP), and image-to-text recall at rank 10 (R@10). Effective protection lowers target performance while retaining non-target performance relative to Clean training; we summarize this separation with Gap, the non-target score minus the target score.

Baselines and models. Clean and uniform random perturbations (Random) serve as reference controls. We compare with EM (Huang et al., 2021), REM (Fu et al., 2022), UC (Zhang et al.,

2023), One for All (Chen et al., 2024), MUE (Liu et al., 2024b), T2UE (Ma et al., 2025), and PIL (Liu et al., 2026) where applicable, under the same ℓ_∞ budget of $8/255$.

4.2 Selective Unlearnability in Image-Text Retrieval

Tables 1 and 2 evaluate image-to-text retrieval with *person* as the Flickr30k target and 20 selected MS COCO classes used as targets in turn. The MS COCO results average over these 20 target choices.

Relative to Clean training, DiffUnlearn lowers target R@10 by 18.73–24.05 points across the four dataset–model settings, while non-target recall falls by only 0.83–2.42 points. This asymmetry shows that the protected images disrupt learning of the chosen concept without similarly degrading retrieval of other concepts. DiffUnlearn has the largest Gap and highest non-target recall among protection methods in all four settings. Some baselines suppress the target more strongly, but their larger non-target losses make their protection less selective; the same pattern on MS COCO shows that the effect extends beyond the Flickr30k *person* target.

Table 1: Selective retrieval on the Flickr30k person-target test set (target and non-target macro R@10, %). Gray marks the three lowest target scores and bold marks the highest non-target and Gap scores among protection methods for each model.

Method	RN50			ViT-B/16		
	Target ↓	Non-target ↑	Gap (pp) ↑	Target ↓	Non-target ↑	Gap (pp) ↑
Clean	40.98	38.70	-2.28	28.33	29.84	1.51
Random	42.13	38.65	-3.48	30.92	29.87	-1.05
One for All	20.22	24.35	4.13	12.22	14.25	2.03
MUE	17.21	15.86	-1.35	7.70	8.24	0.54
T2UE	11.67	23.19	11.52	8.52	13.20	4.68
PIL	32.00	29.02	-2.98	24.27	21.93	-2.34
DiffUnlearn	16.93	**36.28**	**19.35**	8.48	**27.92**	**19.44**

Table 2: Selective image-to-text retrieval on MS COCO val2014 over 20 selected target classes (target and non-target macro R@10, %). Gray marks the three lowest target scores and bold marks the highest non-target and Gap scores among protection methods for each model.

Method	RN50			ViT-B/16		
	Target ↓	Non-target ↑	Gap (pp) ↑	Target ↓	Non-target ↑	Gap (pp) ↑
Clean	34.80	34.34	-0.46	24.29	25.69	1.40
Random	34.75	35.74	0.99	25.93	25.14	-0.79
One for All	17.20	21.25	4.05	10.99	11.90	0.91
MUE	14.23	16.06	1.83	5.97	7.35	1.38
T2UE	11.05	13.44	2.39	4.92	8.16	3.24
PIL	23.65	23.78	0.13	22.39	22.20	-0.19
DiffUnlearn	12.64	**33.51**	**20.87**	5.56	**24.27**	**18.71**

4.3 Transfer across Models and Tasks

Transfer to multilabel classification. Table 3 tests whether non-target concepts remain learnable by a classifier trained on person-protected Flickr30k images. The classifier predicts ten non-target labels, while every training image contains *person* and receives a person-targeted perturbation.

DiffUnlearn achieves 31.04 non-target macro AP, only 0.81 points below Clean and 7.36 points above the strongest competing protection method, T2UE. Thus, the non-target signal survives the change from image–text retrieval to classification even though every training image carries the person-targeted perturbation.

Table 4 tests both target suppression and non-target preservation in multilabel classification over the same 20 MS COCO concepts. Each concept serves as the target in turn, with the other 19

Table 3: Cross-task non-target classification AP (%, ↑) on the person-positive Flickr30k subset. Bold marks the highest protection-method AP per column; italic Clean and Random are unranked reference controls.

Training condition	Dog	Bicycle	Car	Boat	Guitar	Ball	Bag	Chair	Table	Hat	Non-target macro
Clean	18.57	23.27	17.42	16.89	46.41	70.43	23.39	23.33	34.13	44.67	31.85
Random	17.90	23.20	17.44	13.07	46.32	70.06	27.64	26.76	28.04	41.25	31.17
One for All	11.93	13.62	12.46	11.67	26.63	36.44	17.59	15.68	22.62	26.93	19.56
MUE	13.10	17.82	14.27	12.56	27.37	35.03	17.06	16.15	18.05	27.10	19.85
T2UE	16.46	18.80	12.83	11.54	32.18	43.55	20.15	20.04	29.55	31.73	23.68
PIL	13.65	16.80	15.60	12.32	30.79	38.03	17.83	17.28	20.23	26.96	20.95
DIFFUNLEARN	**20.94**	**20.55**	**17.39**	**15.99**	**44.44**	**67.22**	**23.90**	**22.53**	**33.31**	**44.08**	**31.04**

contributing to non-target mAP. For Clean, the averaged target AP equals the averaged non-target mAP by construction, since each of the 20 classes serves once as the target.

Table 4: Cross-task multilabel classification over 20 selected MS COCO classes (AP, %). Each class serves as the target in turn; non-target mAP averages the other 19 classes. Gray marks the two lowest target scores; bold marks the highest non-target and gap scores among protection methods. R101 denotes ResNet-101 and ViT-L denotes ViT-L/16.

Method	**Target AP ↓**				**Non-target mAP ↑**				Gap↑
	Q2L/R101	CSRA/R101	CSRA/ViT-L	Q2L/Swin-L	Q2L/R101	CSRA/R101	CSRA/ViT-L	Q2L/Swin-L	
Clean	84.91	83.26	85.98	90.50	84.91	83.26	85.98	90.50	0.00
Random	85.96	81.89	85.99	90.51	85.13	82.31	85.15	90.78	-0.25
One for All	51.10	49.52	52.63	55.15	65.57	68.31	66.26	61.08	13.21
MUE	69.00	71.71	72.09	80.41	65.98	71.73	77.68	80.42	0.65
T2UE	53.85	52.13	57.90	62.17	62.05	65.21	64.84	63.68	7.43
PIL	48.42	47.68	54.73	55.94	58.70	57.46	59.04	53.83	5.57
DIFFUNLEARN	47.05	48.21	52.84	54.36	**84.22**	**82.51**	**85.32**	**90.40**	**35.00**

Across the four architectures, DIFFUNLEARN reduces mean target AP by 33.14–37.86 points relative to Clean while non-target mAP changes by at most 0.75 points. The target suppression therefore does not come from a broad loss of classification utility. DIFFUNLEARN also retains 7.64–18.24 points more non-target mAP than MUE, the strongest competing protection method on that measure, and achieves a mean Gap of 35.00 points versus 13.21 for the next-best method. The result is consistent across the evaluated classifier families, supporting transfer of selective protection beyond the retrieval objective.

Transfer across visual representations. Table 5 evaluates whether person-targeted protection persists across six frozen visual representations. The non-target result averages dog, horse, car, and bicycle accuracy.

Across the six representations, DIFFUNLEARN reduces mean target accuracy by 15.81 points relative to Clean training while mean non-target accuracy decreases by only 0.36 points. Target accuracy falls in every column, so suppression is not confined to the surrogate representation; the drop ranges from 6.67 points with CLIP to 42.22 points with OpenFlamingo, showing that its strength depends on the representation. Non-target accuracy remains within 1.90 points of Clean in every column, supporting selective transfer under the evaluated frozen-feature setting.

4.4 VISUAL QUALITY

Figure 3 shows grid-like horizontal and vertical bands in T2UE and repeating stripe-like structures in One for All, which remain visible across relatively uniform background regions of the protected images. By contrast, DIFFUNLEARN produces less regular patterns that follow the person's contours and existing scene textures; in these two examples, the perturbations are less conspicuous at the displayed scale while the person and surrounding scene remain recognizable. Its MAE is nevertheless the highest in both examples (6.05 and 6.57 on the 0–255 scale), indicating that visibility here de-

Table 5: Target and non-target accuracy across six frozen visual representations (%). Person is the target; non-target accuracy averages dog, horse, car, and bicycle. Lower target and higher non-target values indicate stronger selective transfer; italic Clean and Random are unranked reference controls.

Method	BLIP-2	CLIP	InstructBLIP	LLaVA	OpenFlamingo	Qwen2.5-VL	Mean
Target accuracy ↓							
Clean	93.33	95.56	93.33	88.64	91.36	92.10	92.39
Random	92.59	94.57	92.59	86.17	89.88	89.63	90.91
DIFFUNLEARN	84.20	88.89	83.95	68.89	49.14	84.44	76.58
Non-target accuracy ↑							
Clean	92.60	95.95	93.24	89.04	91.13	91.37	92.22
Random	92.51	94.69	92.97	89.42	91.98	92.23	92.30
DIFFUNLEARN	92.53	94.05	93.77	89.14	90.11	91.54	91.86

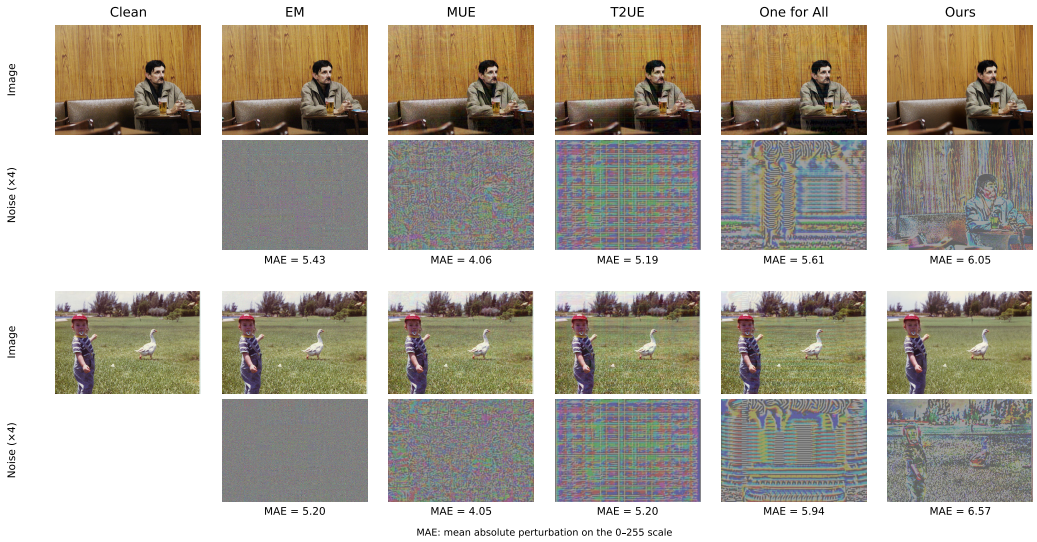

Figure 3: Visual comparison of clean and person-target protected images. Noise panels are amplified fourfold, and the values below show unamplified mean absolute perturbation (MAE) on the 0–255 pixel scale.

pends on the spatial pattern as well as the average magnitude. These examples provide qualitative evidence, not a dataset-wide perceptual-quality ranking.

5 CONCLUSION

This work introduced selective unlearnability in images, allowing data owners to specify a target concept through language while retaining the training utility of other concepts in the same image. DIFFUNLEARN combines diffusion inversion, image-text alignment, target-token attention suppression, and visual preservation under a bounded perturbation to pursue this goal. The completed experiments demonstrate basic unlearnability and provide evidence that protection transfers beyond the surrogate model while largely retaining aggregate non-target utility. Future work will investigate more efficient protection generation, selection of individual object instances through referring expressions, and facial identity protection that preserves other identities and visual information unrelated to identity.

AI USE STATEMENT

In preparing this manuscript, we used ChatGPT to assist with drafting and language editing, checking implementation descriptions against local code, and developing mathematical analyses and proofs. The authors are responsible for verifying the resulting text, mathematical arguments, and correspondence with the experiments, and for the final content of the manuscript.

ETHICS STATEMENT

Our work studies privacy protection for visual concepts using publicly available benchmark datasets. Our evaluation concerns category-level protection and does not establish protection of individual identities.

REPRODUCIBILITY STATEMENT

Implementation details, training hyperparameters, evaluation protocols, and component ablations are provided in Appendix A. Mathematical analyses and proofs are provided in Appendix B. Anonymous code is available at `https://anonymous.4open.science/r/9F6B/`.

REFERENCES

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22560–22570, 2023. URL https://arxiv.org/abs/2304.08465.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270. USENIX Association, 2023. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

Chaochao Chen, Jiaming Zhang, Yuyuan Li, and Zhongxuan Han. One for all: A universal generator for concept unlearnability via multi-modal alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7700–7711, 2024. URL https://proceedings.mlr.press/v235/chen24bc.html.

Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. Diffusion models for imperceptible and transferable adversarial attack. *arXiv preprint arXiv:2305.08192*, 2023. doi: 10.48550/arXiv.2305.08192. URL https://arxiv.org/abs/2305.08192.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8717–8730. PMLR, 2023. URL https://proceedings.mlr.press/v202/duan23b.html.

Shaopeng Fu, Fengxiang He, Yang Liu, Li Shen, and Dacheng Tao. Robust unlearnable examples: Protecting data privacy against adversarial learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=baUQQPwQiAg.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_CDixzkzeyb.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=iAmZUo0DxC0.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

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, 2014.

Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7817–7826, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Towards_Understanding_Cross_and_Self-Attention_in_Stable_Diffusion_for_Text-Guided_CVPR_2024_paper.html.

Jinlin Liu, Wei Chen, and Xiaojin Zhang. Perturbation-induced linearization: Constructing unlearnable data with solely linear classifiers. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GBSGToE97J.

Xinwei Liu, Xiaojun Jia, Yuan Xun, Siyuan Liang, and Xiaochun Cao. Multimodal unlearnable examples: Protecting data against multimodal contrastive learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8024–8033, 2024b. doi: 10.1145/3664647.3680708. URL https://arxiv.org/abs/2407.16307.

Xingjun Ma, Hanxun Huang, Tianwei Song, Ye Sun, Yifeng Gao, and Yu-Gang Jiang. T2UE: Generating unlearnable examples from text descriptions. *arXiv preprint arXiv:2508.03091*, 2025. URL https://arxiv.org/abs/2508.03091.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. NULL-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6038–6047, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Mokady_NULL-Text_Inversion_for_Editing_Real_Images_Using_Guided_Diffusion_Models_CVPR_2023_paper.html.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Jie Ren, Han Xu, Yuxuan Wan, Xingjun Ma, Lichao Sun, and Jiliang Tang. Transferable unlearnable examples. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=-htnolWDLvP.

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*, pp. 10684–10695, 2022.

Vinu Sankar Sadasivan, Mahdi Soltanolkotabi, and Soheil Feizi. CUDA: Convolution-based unlearnable datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3862–3871, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Sadasivan_CUDA_Convolution-Based_Unlearnable_Datasets_CVPR_2023_paper.html.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: Protecting artists from style mimicry by Text-to-Image models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2187–2204. USENIX Association, 2023. URL https://www.usenix.org/conference/usenixsecurity23/presentation/shan.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Somepalli_Diffusion_Art_or_Digital_Forgery_Investigating_Data_Replication_in_Diffusion_CVPR_2023_paper.html.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Weixiang Sun, Yixin Liu, Zhiling Yan, Kaidi Xu, and Lichao Sun. Medical unlearnable examples: Securing medical data from unauthorized training via sparsity-aware local masking. *arXiv preprint arXiv:2403.10573*, 2024. URL https://arxiv.org/abs/2403.10573.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.310. URL https://aclanthology.org/2023.acl-long.310/.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1921–1930, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Tumanyan_Plug-and-Play_Diffusion_Features_for_Text-Driven_Image-to-Image_Translation_CVPR_2023_paper.html.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N. Tran, and Anh Tran. Anti-DreamBooth: Protecting users from personalized Text-to-Image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2116–2127, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Van_Le_Anti-DreamBooth_Protecting_Users_from_Personalized_Text-to-image_Synthesis_ICCV_2023_paper.html.

Jiaming Zhang, Xingjun Ma, Qi Yi, Jitao Sang, Yu-Gang Jiang, Yaowei Wang, and Changsheng Xu. Unlearnable clusters: Towards label-agnostic unlearnable examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3984–3993, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Unlearnable_Clusters_Towards_Label-Agnostic_Unlearnable_Examples_CVPR_2023_paper.html.

Chengshuai Zhao, Zhen Tan, Dawei Li, Zhiyuan Yu, and Huan Liu. To see is not to learn: Protecting multimodal data from unauthorized fine-tuning of large vision-language model. *arXiv preprint arXiv:2605.14291*, 2026. URL https://arxiv.org/abs/2605.14291.

A Supplementary Experimental Results

This appendix provides implementation details, supplementary classification results, component ablations, and mathematical analysis.

A.1 Implementation Details

The main Flickr30k protection set is generated at 256×256 resolution with 60 latent updates and latent optimizer learning rate 0.02. Unless otherwise stated, all reported quantitative results are averaged over three random seeds. Every decoded update is projected into the ℓ_∞ ball of radius $\epsilon = 8/255$ around the clean image. Across the reported experiments, VLP alignment uses only the target text s, minimizing its cosine distance to the protected image in the VLP embedding space. Surrogate training also minimizes cosine distance to the target text, updating both CLIP encoders; original image captions are not used during protection generation. The target text may use a fixed template, such as `a photo of a person` for `person`. Special tokens are handled by the pretrained tokenizer, with no separately trained special-token embeddings.

The six representation columns in Table 5 use BLIP-2 Flan-T5-XL, CLIP ViT-L/14-336, InstructBLIP, LLaVA-1.5-7B, OpenFlamingo vision, and Qwen2.5-VL-7B, respectively.

Reference generation configuration. The reference implementation uses Stable Diffusion 2-base and OpenAI CLIP ViT-B/16. The visual-loss normalization uses $\zeta = 10^{-6}$; loss weights are specified below. Latent optimization uses Adam with learning rate 0.02. The diffusion model and VAE remain frozen.

The scheduler uses a 20-step DDIM leading timetable, epsilon prediction, and deterministic sampling ($\eta = 0$). Inversion uses empty-text conditioning with guidance scale 0; subsequent denoising uses the target prompt with guidance scale 2.5. Cross-attention is aggregated over the down, middle, and up blocks, averaging heads and target tokens within each layer and averaging over layers and steps. The reference mask is computed once, normalized, and detached. The diffusion prompt template is `a photo of a {target}`. Unless otherwise noted, all experiments use these settings; any experiment-specific deviations are described in the corresponding experimental setup.

Loss-scale calibration and weight selection. The three losses have different natural scales: cosine alignment lies in $[0, 2]$, attention suppression averages target-token probabilities, and visual preservation measures squared pixel differences under a small perturbation budget. Equal coefficients therefore need not yield comparable numerical contributions. To calibrate their relative weights for a dataset–surrogate configuration, evaluate the unweighted losses on a fixed calibration subset of the protection training data at the initialized, decoded, and projected candidates, before the first latent update. Let $\ell_i^{(0)} > 0$ be the mean initial value for $i \in \{\mathrm{align}, \mathrm{attn}, \mathrm{vis}\}$. Fixing the alignment weight to one, we set

$$\lambda_{\mathrm{align}} = 1, \qquad \lambda_{\mathrm{attn}} = \frac{\ell_{\mathrm{align}}^{(0)}}{\ell_{\mathrm{attn}}^{(0)}}, \qquad \lambda_{\mathrm{vis}} = \frac{\ell_{\mathrm{align}}^{(0)}}{\ell_{\mathrm{vis}}^{(0)}}. \tag{9}$$

The three weighted initial losses therefore share the same scale, $\ell_{\mathrm{align}}^{(0)}$.

Calibration is performed separately for each dataset–surrogate configuration, and the resulting weights remain fixed during subsequent alternating optimization. In particular, replacing cosine alignment with a classification loss requires fresh calibration. This rule balances initial loss values; it does not enforce equal latent-gradient norms or maintain equal contributions throughout training.

Downstream training hyperparameters. For image–text retrieval on Flickr30k and MS COCO, RN50 and ViT-B/16 are trained for 32 epochs using AdamW with an initial learning rate of 5×10^{-4}. The Flickr30k non-target classifier (Q2L/Swin-L) is trained for 20 epochs using AdamW with an initial learning rate of 10^{-4}. All four MS COCO classifiers (Q2L/R101, CSRA/R101, CSRA/ViT-L, and Q2L/Swin-L) are trained for 20 epochs using SGD with an initial learning rate of 10^{-2}. For CIFAR-10 and CIFAR-100, all four classifiers are trained for 60 epochs using SGD with an initial learning rate of 0.1. For frozen-representation transfer, each classifier head is trained for up to 1,000 optimization steps using AdamW with a learning rate of 10^{-2}.

Flickr30k non-target classification subset. For the ten-label classification evaluation in Table 3, images are retained when at least one caption mentions person (including matched person-related words) and at least one of dog, bicycle, car, boat, guitar, ball, bag, chair, table, or hat. Labels are

derived from caption vocabulary matches; an unmentioned concept is treated as a weak negative, not a verified visual absence. The source train/validation/test assignment is preserved, with no filename overlap between splits. Person determines the protection target and is excluded from the ten classifier outputs. Validation and test images remain clean. The retrieval use of these images and their captions is specified separately below.

Classification evaluation and random control. Classification is evaluated on clean held-out images. For Flickr30k, per-label AP is computed using an interpolated precision envelope and macro-averaged across the ten non-target outputs. CIFAR evaluation reports clean-test accuracy. Random adds one fixed per-image uniform perturbation within the same ℓ_∞ budget before preprocessing.

Retrieval data and target selection. Flickr30k retrieval reuses the person co-occurrence splits above and restores each image's original captions. For COCO retrieval, the 20 targets follow the first 20 entries of the saved index ordered by official category ID; they are not selected by frequency. The targets are person, bicycle, car, motorcycle, airplane, bus, train, truck, boat, traffic light, fire hydrant, stop sign, parking meter, bench, bird, cat, dog, horse, sheep, and cow. Each target uses its target-positive train2014 cohort for training and clean val2014 cohort for evaluation. For each target, candidate non-target categories are selected using only its training cohort, requiring at least 30 positive and 30 negative images per category. The number of such categories varies by target. This retrieval protocol differs from the 20-output classification evaluation, where non-target mAP averages the other 19 outputs.

Concept-specific image-to-text recall. The candidate gallery contains all original captions of the images in the current evaluation cohort. For an image query and concept c, a relevant caption must both belong to that query image and mention c according to the fixed concept vocabulary; the image must also carry the concept label. Queries without any such caption are excluded from that concept's denominator. A hit at rank K occurs when at least one relevant caption appears among the top K. An image or caption may contribute to multiple concept evaluations when concepts co-occur; the target and non-target query sets are not assumed disjoint. Non-target recall is macro-averaged over categories with eligible queries. COCO scores are then averaged equally over the 20 targets, rather than weighted by cohort size. Query sets and galleries are held fixed across compared conditions.

Retrieval evaluation. We evaluate RN50 and ViT-B/16 using normalized embeddings and cosine similarity on clean images and the unmodified caption gallery.

Frozen-representation transfer. For Table 5, the pretrained visual feature extractors remain frozen, and a separate classifier is fitted for each training condition. Person is the target; dog, horse, car, and bicycle are the non-target classes. This single-label probe retains caption-derived records matching exactly one of these five classes and preserves the original Flickr30k splits. A class-balanced training subset is sampled; all eligible validation and test images remain clean. Only target-class training images are replaced by their protected versions at budget $8/255$. This probe measures transfer across representations and is distinct from the co-occurring multilabel utility evaluation.

Extracted features are ℓ_2-normalized. CLIP uses its image embedding; BLIP-2 and InstructBLIP use mean patch-token features excluding the class token; LLaVA and Qwen2.5-VL use mean image-token features; the OpenFlamingo column uses its frozen OpenCLIP ViT-L/14 visual tower and mean patch-token features. The corresponding checkpoints are BLIP-2 Flan-T5-XL, CLIP ViT-L/14-336, InstructBLIP Flan-T5-XL, LLaVA-1.5-7B, Qwen2.5-VL-7B-Instruct, and OpenAI-pretrained ViT-L/14, respectively. Qwen's image processor caps the pixel count at 256×256; other processors follow their model-specific image preprocessing.

Non-target accuracy averages the four non-target classes.

A.2 Basic Unlearnability in Single-Label Classification

On CIFAR-10 and CIFAR-100, DiffUnlearn reduces clean-test accuracy to 10.40–18.24% and 2.42–3.72%, respectively, across four classifier architectures. Table 6 provides the full comparison: these accuracies are far below the corresponding Clean ranges of 89.09–94.27% and 64.37–76.06%.

Table 6: Clean-test accuracy (%, ↓) on CIFAR-10 and CIFAR-100. Gray shading marks the three lowest protection-method accuracies per row; italic Clean and Random are unranked reference controls.

Dataset	Model	*Clean*	*Random*	EM	REM	UC	One for All	MUE	T2UE	PIL	Ours
CIFAR-10	ResNet-18	94.27	93.31	19.71	25.82	79.71	39.76	90.28	13.90	13.07	10.40
	ResNet-50	93.94	91.08	16.79	20.70	75.57	39.00	89.44	10.68	19.40	18.24
	VGG-16	93.27	92.70	16.28	25.09	84.47	38.43	88.84	22.57	15.18	16.13
	DenseNet-121	89.09	88.90	18.58	24.10	85.32	37.34	85.85	13.25	17.35	15.81
CIFAR-100	ResNet-18	75.59	72.62	13.96	10.21	44.35	22.03	72.04	1.93	2.72	2.98
	ResNet-50	76.06	73.01	10.85	9.83	37.40	23.26	73.50	1.65	2.41	3.72
	VGG-16	71.02	68.49	5.83	9.58	54.96	21.83	67.01	11.87	2.42	2.51
	DenseNet-121	64.37	64.19	11.66	13.20	52.25	19.51	59.20	3.34	2.29	2.42

This confirms that the method also impairs ordinary single-label learning beyond the surrogate used to construct the images. DIFFUNLEARN ranks among the three strongest protection methods in all eight settings and is best on CIFAR-10 with ResNet-18. T2UE or PIL achieves lower accuracy in the other settings, but DIFFUNLEARN remains competitive despite also aiming to preserve non-target concepts in the selective experiments.

A.3 COMPONENT ABLATION

Table 7 evaluates the contribution of each loss to Flickr30k person-target protection on RN50 and ViT-B/16. Each ablation removes one loss while keeping the image set, initialization, perturbation budget, remaining loss weights, latent-update count, and downstream training protocol fixed. Removing attention suppression retains the reference mask in the visual loss; removing visual preservation retains the hard ℓ_∞ constraint. We report target R@10, non-target macro R@10, and Gap, defined as non-target minus target R@10 in percentage points.

The full method achieves the largest Gap on both architectures: 19.35 pp on RN50 and 19.44 pp on ViT-B/16. Removing alignment increases target R@10 from 16.93% to 31.57% and from 8.48% to 18.61%, respectively, indicating weaker target suppression. Removing attention suppression also increases target R@10, to 22.64% and 14.63%, while reducing non-target recall on both architectures. These results support the contribution of both target-guidance terms to selective protection.

Removing visual preservation further lowers target R@10 to 15.29% and 7.17%, but also reduces non-target R@10 from 36.28% to 30.56% and from 27.92% to 22.97%, respectively. Consequently, Gap falls to 15.27 pp and 15.80 pp. This trade-off supports the role of visual preservation in retaining non-target utility, even with the hard perturbation constraint in place.

Table 7: Component ablations for Flickr30k person protection at $\epsilon = 8/255$ (target and non-target macro R@10, %). Each loss ablation sets its corresponding weight to zero. Clean and full-method retrieval scores are reproduced from Table 1. Gap is non-target minus target R@10 in percentage points.

Configuration	RN50			ViT-B/16		
	Target ↓	Non-target ↑	Gap (pp) ↑	Target ↓	Non-target ↑	Gap (pp) ↑
Clean	40.98	38.70	-2.28	28.33	29.84	1.51
Full DIFFUNLEARN	16.93	36.28	19.35	8.48	27.92	19.44
No alignment	31.57	35.39	3.82	18.61	25.45	6.84
No attention loss	22.64	31.96	9.32	14.63	24.59	9.96
No visual loss	15.29	30.56	15.27	7.17	22.97	15.80

A.4 Decoding and Gradient Computation

Let $\mathcal{R}_\theta(z_{i,\tau}, p_i)$ denote deterministic denoising from step τ to timestep zero. At iteration k, decoding gives a decoded candidate image

$$\widehat{x}_i^{(k)} = \mathcal{G}\Big(\mathcal{R}_\theta(z_{i,\tau}^{(k)}, p_i)\Big). \tag{10}$$

We project it onto the intersection of the valid image range and the ℓ_∞ ball around x_i:

$$\Pi_{\mathcal{X}_\epsilon(x_i)}(\widehat{x}) = \mathrm{clip}_{[0,1]}\Big(x_i + \mathrm{clip}_{[-\epsilon,\epsilon]}(\widehat{x} - x_i)\Big). \tag{11}$$

Applying this operator to the decoded candidate image at iteration k gives the protected candidate image

$$\widetilde{x}_i^{(k)} = \Pi_{\mathcal{X}_\epsilon(x_i)}(\widehat{x}_i^{(k)}). \tag{12}$$

To avoid zero gradients in saturated clipping regions, we use a straight-through estimator: the forward pass uses the protected candidate image, while the backward pass approximates the projection Jacobian by the identity. With $\mathrm{sg}(\cdot)$ denoting stop-gradient, this is written as

$$\widetilde{x}_{i,\mathrm{st}}^{(k)} = \widehat{x}_i^{(k)} + \mathrm{sg}\Big(\widetilde{x}_i^{(k)} - \widehat{x}_i^{(k)}\Big). \tag{13}$$

The alignment and visual preservation terms are evaluated on the protected candidate image using Equation 13, whereas the attention term is evaluated along the denoising trajectory before pixel projection. With nonnegative weights λ_{align}, λ_{attn}, and λ_{vis}, the combined loss at iteration k is

$$\begin{aligned}\mathcal{L}_{\mathrm{sel}}^{(k)} = {} & \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}}(\widetilde{x}_{i,\mathrm{st}}^{(k)}; s) \\ & + \lambda_{\mathrm{attn}}\mathcal{L}_{\mathrm{attn}}(a_i^{(k)}; M_i) \\ & + \lambda_{\mathrm{vis}}\mathcal{L}_{\mathrm{vis}}(\widetilde{x}_{i,\mathrm{st}}^{(k)}; x_i, M_i).\end{aligned} \tag{14}$$

B Mathematical Feasibility and Conditional Analysis

We distinguish properties guaranteed by the construction from properties requiring assumptions about representations and downstream learning. The projection and distortion bounds below follow from the stated method. The local-direction result gives a sufficient condition for compatible guidance objectives. A separate population regression example shows that bounded, target-positive-only perturbations can impair target learning without impairing an independent non-target task. These results do not guarantee selective unlearnability for arbitrary neural learners, nor do they establish the reported AP or retrieval scores.

B.1 Feasibility of the Released Images

Proposition 1 (exact constraint satisfaction). For any $x \in [0,1]^d$, $\epsilon \geq 0$, and decoded candidate $\widehat{x} \in \mathbb{R}^d$, the map in Equation 11 is the Euclidean projection onto

$$\mathcal{X}_\epsilon(x) = \{y \in [0,1]^d : \|y - x\|_\infty \leq \epsilon\}. \tag{15}$$

In particular, every projected candidate satisfies the release constraint, regardless of the latent update or its convergence.

Proof. The feasible set is the nonempty box $\prod_{j=1}^d [l_j, u_j]$, where $l_j = \max(0, x_j - \epsilon)$ and $u_j = \min(1, x_j + \epsilon)$; it contains x. Minimizing $\sum_j (y_j - \widehat{x}_j)^2$ over this box separates by coordinate and has the unique solution $y_j = \min(u_j, \max(l_j, \widehat{x}_j))$. Since $x_j \in [0,1]$, clipping first to $[x_j - \epsilon, x_j + \epsilon]$ and then to $[0,1]$ gives the same expression. Thus Equation 11 is this projection. The straight-through estimator changes only the backward computation and not this forward guarantee. □

This proposition applies to the floating-point projected image. Any subsequent quantization or resizing must be checked separately against the budget in the released representation.

B.2 What Visual Preservation Controls

Proposition 2 (distortion in low-attention regions). Fix a reference spatial mask M and let $\overline{M} \in [0,1]^d$ denote its broadcast over scalar pixel coordinates. Write $D = \|1 - M\|_1 + \zeta$ for the spatial-mask denominator in Equation 7. For $0 < \kappa \leq 1$, let $N_\kappa = \{j : 1 - \overline{M}_j \geq \kappa\}$. If the projected candidate has $\mathcal{L}_{\text{vis}} \leq v$, then

$$\|\delta\|_2^2 \leq dv, \qquad \|\delta_{N_\kappa}\|_2^2 \leq \min\left\{dv, \frac{Dv}{\kappa^2}, |N_\kappa|\epsilon^2\right\}. \tag{16}$$

Proof. Both terms of $\mathcal{L}_{\text{vis}}$ are nonnegative, so its second term implies $\|\delta\|_2^2 \leq dv$. On N_κ, $(1 - \overline{M}_j)^2 \geq \kappa^2$, hence

$$\kappa^2 \sum_{j \in N_\kappa} \delta_j^2 \leq \sum_j (1 - \overline{M}_j)^2 \delta_j^2 \leq D\mathcal{L}_{\text{vis}} \leq Dv. \tag{17}$$

The hard budget also gives $\sum_{j \in N_\kappa} \delta_j^2 \leq |N_\kappa|\epsilon^2$. Taking the minimum proves the claim. □

For any fixed L-Lipschitz representation h depending only on these coordinates, this yields $\|h(\widetilde{x}) - h(x)\|_2 \leq L\|\delta_{N_\kappa}\|_2$. This is a conditional feature-stability statement: low target attention need not identify every non-target concept, and stability of a fixed representation alone does not prove preservation after retraining.

B.3 Compatibility of the Two Target Guidance Terms

Proposition 3 (a sufficient local condition). During a latent update, fix the surrogate, diffusion model, and reference mask. Let z collect the optimized latents, $f_1(z) = \mathcal{L}_{\text{align}}$, and $f_2(z) = \mathcal{L}_{\text{attn}}$, using the actual forward computation. Let $h(z)$ be a chosen vector of non-target features of the projected images. Assume that these functions are twice continuously differentiable near z_0. Write

$$B = J_h(z_0), \quad P = \text{proj}_{\ker B}, \quad G = [\nabla f_1(z_0), \nabla f_2(z_0)], \quad H = G^\top P G. \tag{18}$$

If H is positive definite, there is a latent direction q that decreases both guidance losses to first order while leaving h unchanged to first order. Specifically, with $\mathbf{1} = (1,1)^\top$,

$$q = -PGH^{-1}\mathbf{1}, \qquad f_j(z_0 + tq) = f_j(z_0) - t + O(t^2), \qquad h(z_0 + tq) = h(z_0) + O(t^2). \tag{19}$$

Proof. Because P projects onto $\ker B$, $BP = 0$, and therefore $Bq = 0$. Also $G^\top q = -G^\top PGH^{-1}\mathbf{1} = -\mathbf{1}$. Taylor expansion along q gives both statements. For sufficiently small positive t, the linear loss reductions dominate the quadratic remainders. All resulting projected images still satisfy Proposition 1. □

The rank condition means that two independent target-guidance directions remain after restricting to first-order non-target-preserving directions; it is a condition to verify, not an asserted property of every image. Differentiability may be considered within a fixed clipping region, away from its boundary. The result uses true forward derivatives. It does not prove that the implemented Adam update with straight-through gradients selects q, that each update decreases both losses, or that alternating surrogate training converges.

B.4 A Constructive Example of Selective Learning Degradation

Proposition 4 (target-positive shortcuts under ridge regression). Let $Y \in \{-1, 1\}$ be equiprobable, and let ξ be independent of Y, with zero mean, variance σ^2, and $|\xi| \leq h$. Choose $a > 0$ with $a + h \leq 1/2$. Let (U, T) be independent of (Y, ξ), where $U \in [0,1]^m$ represents non-target content and $\mathbb{E}[T^2] < \infty$. The clean input is

$$X = (1/2 + aY + \xi,\ 1/2,\ U). \tag{20}$$

For any $\epsilon > 0$ and $0 < b \leq \min\{\epsilon, 1/2\}$, release

$$\widetilde{X} = X + (0,\ b\,\mathbf{1}_{\{Y=1\}},\ 0). \tag{21}$$

Train separate population ridge regressors for Y and T, using all input coordinates, an unpenalized intercept, and a squared-weight penalty $\lambda > 0$. Then: (i) only target-positive examples are perturbed and the release constraint is satisfied; (ii) the target regressor's clean-test squared error strictly increases compared with clean training; and (iii) the non-target regressor and its clean-test error are unchanged.

Proof. The first statement follows immediately from $0 < b \le \min\{\epsilon, 1/2\}$ and $X \in [0,1]^{m+2}$. For the target task, write $R = aY + \xi$, $S = b\mathbf{1}_{\{Y=1\}}$, $V = U - \mathbb{E}[U]$, and $c = b/2$. Absorbing the fixed coordinate offsets into the intercept, the regression objective is

$$\mathbb{E}\big[(Y - \beta - w_r R - w_s S - w_u^\top V)^2\big] + \lambda(w_r^2 + w_s^2 + \|w_u\|_2^2). \tag{22}$$

Independence implies $w_u = 0$. The optimal intercept is $\beta = -cw_s$, and $S - c = cY$. Thus the centered normal equations are

$$\begin{bmatrix} a^2 + \sigma^2 + \lambda & ac \\ ac & c^2 + \lambda \end{bmatrix} \begin{bmatrix} w_r \\ w_s \end{bmatrix} = \begin{bmatrix} a \\ c \end{bmatrix}. \tag{23}$$

Their unique solution has

$$\begin{aligned} D_b &= \lambda(a^2 + \sigma^2 + \lambda) + c^2(\sigma^2 + \lambda), \\ w_r(b) &= \frac{a\lambda}{D_b}, \qquad w_s(b) = \frac{c(\sigma^2 + \lambda)}{D_b}. \end{aligned} \tag{24}$$

Clean training gives $w_r(0) = a/(a^2 + \sigma^2 + \lambda)$ and $w_s(0) = \beta(0) = 0$. Since $b > 0$, $0 < w_r(b) < w_r(0)$. On clean test inputs $S = 0$, so the target risk is

$$\mathcal{R}_Y(b) = 1 - 2aw_r(b) + (a^2 + \sigma^2)w_r(b)^2 + c^2 w_s(b)^2. \tag{25}$$

The function $q(w) = 1 - 2aw + (a^2 + \sigma^2)w^2$ is strictly decreasing on $[0, w_r(0)]$, because $(a^2 + \sigma^2)w_r(0) < a$. Therefore $q(w_r(b)) > q(w_r(0))$, and the nonnegative intercept-error term implies $\mathcal{R}_Y(b) > \mathcal{R}_Y(0)$.

For the non-target task, independence gives $\mathrm{Cov}(T, R) = \mathrm{Cov}(T, S) = 0$, and the covariance between V and (R, S) is zero. The ridge normal equations therefore assign zero weights to R, S in both conditions. The remaining solution is

$$w_u = (\mathrm{Cov}(U) + \lambda I)^{-1} \mathrm{Cov}(U, T), \tag{26}$$

with prediction $\mathbb{E}[T] + w_u^\top(U - \mathbb{E}[U])$, unchanged by the perturbation. This proves (iii). □

This example establishes the existence of selective degradation under a specified learning rule and data distribution, even when only target-positive samples are modified. The perturbation is constructed analytically; the proposition does not show that diffusion latent optimization discovers it. Independence of target and non-target factors and the population ridge objective are explicit idealizations. Together with Proposition 3, the example motivates searching for accessible target-correlated cues with limited non-target interference. Whether those conditions hold for real images and whether the effect transfers to AP or R@10 remain empirical questions.