Beyond Feature Disruption: Boundary-Diverting Unlearnable Examples against Linear Probing

Zhihao Li[1] **Jiale Cai**[1] **Gezheng Xu**[1,†] **Ruiyi Fang**[1] **Hao Zheng**[2]
Ruizhi Pu[3] **Zhong Ji**[4] **Charles Ling**[1] **Boyu Wang**[1,5]
[1] Western University [2] Shenzhen Campus of Sun Yat-sen University
[3] Southeast University [4] Tianjin University [5] Vector Institute
{zli3446,jcai336,gxu86,rfang32,charles.ling}@uwo.ca
zhengh239@mail.sysu.edu.cn ruizhipu@seu.edu.cn jizhong@tju.edu.cn
bwang@csd.uwo.ca

Abstract

Unlearnable Examples (UEs) have emerged as a promising data protection strategy against unauthorized model training, which adds imperceptible perturbations into data to degrade generalization. Existing studies predominantly assume a fully trainable target model, where unlearnability is achieved by disrupting clean feature representations. However, this assumption grows increasingly unrealistic with the prevalence of pretrained models, where unauthorized users can freeze the backbone and perform linear probing, preserving discriminative representations and thereby invalidating the core mechanism of previous UEs. In this paper, we investigate this practical yet challenging linear probing setting and reveal a fundamental vulnerability. To address this, we propose **DIVERT** (**D**ecision-boundary d**I**version **V**ia s**E**mantic-decoupled o**R**thogonal **T**argets), a novel framework that shifts the design principle from feature disruption to explicit decision boundary diversion. Our intuition is that projecting perturbed samples into a subspace orthogonal to the semantic manifold induces spurious linear separability, thereby steering the decision boundary away from original prototypes. Building on this principle, DIVERT first constructs synthetic class anchors within the semantic-decoupled subspace, then employs a cross-attention mechanism to jointly optimize these anchors while pushing perturbed samples toward them to establish spurious separability. Subsequently, it incorporates a surrogate linear head to simulate the linear probing procedure, further refining perturbations to actively divert the decision boundary. Extensive experiments demonstrate that DIVERT establishes effective unlearnability across diverse datasets and backbones. Code is available at https://github.com/zhli-cs/DIVERT.

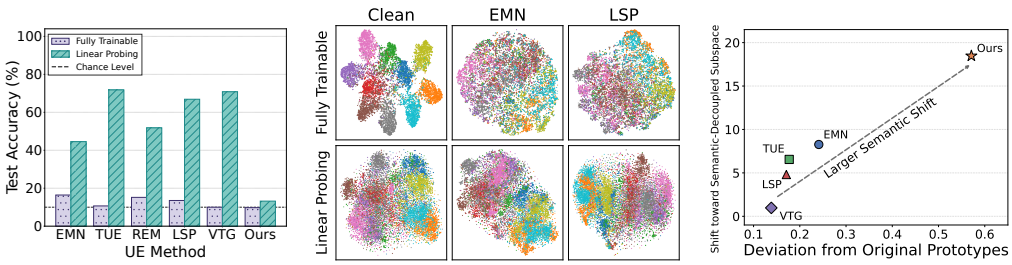
1 Introduction

The rise of the internet has advanced the construction of large-scale datasets [1], greatly accelerating the development of Deep Neural Networks (DNNs) and data-centric deep learning [2, 3, 4, 5, 6]. However, growing concerns are arising regarding unauthorized data usage. To safeguard data privacy, Unlearnable Examples (UEs) [7, 8, 9, 10, 11] have emerged as a promising data protection strategy. By adding imperceptible perturbations to training data, UEs mislead models into learning spurious correlations rather than meaningful semantics, thereby substantially reducing clean test performance.

Existing studies predominantly focus on generating UEs to trick a fully trainable target model. However, with the growing prevalence of pretrained models, practitioners increasingly leverage

[†]Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Fragile Unlearnability (b) Linear Probing Challenge (c) Semantic Shift

Figure 1: Empirical analyses on how UEs behave under linear probing (ResNet-18 on CIFAR-10). (a) Existing UEs perform well in fully trainable settings but fail under linear probing. (b) t-SNE of clean test features (full illustration in Appendix A.1). Unlike full training (top) where UEs disrupt representations, the frozen backbone in linear probing (bottom) nullifies this mechanism and preserves discriminative features. (c) Quantification of perturbation-induced semantic shift along two dimensions (details in Appendix A.2): deviation from original prototypes and shift toward the semantic-decoupled subspace, where our method attains the largest values on both.

pretrained backbones and perform only linear probing [12, 13, 14]. This paradigm maintains satisfactory performance and avoids heavy finetuning on downstream datasets, exhibiting great potential in practical scenarios. However, the efficacy of UEs under this scenario has not been investigated, leaving a critical gap in examining their effectiveness in real-world deployment.

In this paper, we conduct the first UE study under linear probing and reveal a fundamental vulnerability. As illustrated in Figure 1a, existing UEs suffer severe unlearnability degradation when an unauthorized user trains a linear head on a frozen pretrained backbone. We further investigate this phenomenon by visualizing feature representations in Fig. 1b and discover that conventional UEs typically achieve unlearnability by injecting perturbations that corrupt the representation learning process. Nevertheless, linear probing freezes the pretrained backbone and thus is immune to such disruption, naturally invalidating this core mechanism while continuing to produce discriminative features.

To render UEs effective against linear probing, we argue that the design principle must shift from feature disruption to decision boundary diversion. Under this paradigm, even if the backbone generates reliable class clusters, perturbations trick the linear head into fitting displaced decision boundaries. Consequently, the linear head fails to generalize to clean test data, thereby preserving unlearnability. Empirically, linear probing still yields high clean test accuracy after being trained on existing UEs (Figure 1a), which implies that prior methods fail to induce sufficient boundary discrepancy. To quantitatively verify this deficiency, we leverage the perturbation-induced semantic shift as an implicit proxy for boundary diversion, given that decision boundaries are inherently governed by the geometry of class prototypes. As shown in Figure 1c, existing UEs induce only minimal deviation from original prototypes and struggle to project perturbed samples into a semantically decoupled subspace, which motivates us to design UEs explicitly tailored for decision boundary diversion.

To this end, we propose DIVERT, a novel method that enforces unlearnability against linear probing. The core idea is to explicitly divert decision boundaries away from genuine semantics. We achieve this through a two-step optimization framework that first shifts perturbed samples into a semantic-decoupled subspace and then diverts the decision boundary accordingly. In the first step, we construct synthetic class anchors within the semantic-decoupled subspace. We then optimize perturbations to push samples toward these targets, inducing spurious separability and laying the foundation for subsequent boundary diversion. In the second step, to further refine perturbations in actively tricking models into learning the spurious separability, we incorporate a surrogate linear head to simulate the linear probing procedure. Through this two-step process, DIVERT effectively establishes decision boundary discrepancies between perturbed training data and clean test data, thereby reducing test accuracy and injecting strong unlearnability. Our main contributions are summarized as follows:

- We reveal a fundamental vulnerability of existing UEs under linear probing, where the frozen backbone inherently resists perturbation manipulation and invalidates their core mechanism.
- We propose DIVERT, a novel two-step framework that progressively diverts decision boundaries away from genuine semantics. The first step induces spurious separability within a

semantic-decoupled subspace, while the second step anticipates the linear probing procedure to actively steer the decision boundary onto this synthetic discriminability.

- Extensive experiments demonstrate that DIVERT achieves superior unlearnability, with strong transferability across pretraining regimes, datasets, and architectures, and further generalizes to the conventional fully trainable setting.

2 Related Work

As machine learning systems are increasingly deployed in real-world applications, their societal implications have attracted growing research attention, spanning critical dimensions such as privacy [15, 16], safety [17], and fairness [18], and imbalanced learning [19]. Among privacy-preserving machine learning studies [20, 21, 22, 23, 24], Unlearnable Examples [25, 26, 27, 28, 29, 30, 31] have emerged as a promising strategy for safeguarding data against unauthorized exploitation, which adds perturbations to training data such that models trained on such protected data fail to generalize to clean test data. To achieve this, mainstream approaches inject "shortcuts", tricking models to rely on spurious perturbation-label associations rather than acquiring genuine semantics [32, 33, 34, 35]. Among UE studies, a pioneering method is Error-Minimization Noises [7], which formulates a min-min optimization that alternately updates perturbations and models. Building on this foundation, subsequent works have expanded UEs along multiple directions. Several methods [32, 36, 37] extend UEs to unsupervised learning. UC [38] and One-for-All [39] address label-agnostic scenarios, where various labels may be assigned to the same image under different categorization criteria. A growing line of research broadens the scope beyond image classification to other tasks such as segmentation [40], 3D point clouds [9, 41] and object detection [42]. Recent efforts further extend UEs to other modalities, including text [43], voice data [44] and EEG signals [45].

Despite these progresses, existing methods predominantly assume a fully trainable threat model, overlooking the increasingly prevalent linear probing paradigm. In this paper, to the best of our knowledge, we present the first study to uncover the vulnerability of UEs under this regime. We then propose a novel framework that induces spurious linear separability and diverts the decision boundary away from true semantics, successfully rendering data unlearnable under linear probing.

3 Methodology

In this section, we first formalize a novel UE evaluation setting that employs linear probing with a pretrained backbone. We then propose DIVERT, a two-step optimization framework that diverts the decision boundary toward the semantic-decoupled subspace to achieve unlearnability.

3.1 Problem Statement

Unlike existing UE studies that assume a fully trainable target model, we introduce a more realistic setting where unauthorized users leverage a pretrained backbone and perform only linear probing. In this scenario, model backbones remain fixed during training, making the perturbations unable to disrupt internal representations and thereby invalidating the core mechanism of previous UE methods. We formalize the definition of UEs and the new evaluation scenario below.

Threat Model. Unlearnable examples is a data protection strategy designed to prevent unauthorized data exploitation. We assume the *Defender* has full access to the data they wish to protect and may apply any perturbation scheme to make it unlearnable, subject to a bounded perturbation budget. Once the perturbed data is released, the defender cannot intervene in the training process. The *Attacker* is an unauthorized party who scrapes the perturbed data to train DNNs. The attacker has full access to the perturbed dataset but possesses no knowledge of the original clean data, the perturbations, or the perturbation generation procedure. Our proposed method operates on the defender side, aiming to safeguard data against unauthorized exploitation.

Notations and Objective. We formulate the problem in the context of image classification via linear probing. Let $(\boldsymbol{x}, y)$ be a labeled example, where $\boldsymbol{x} \in \mathcal{X}$ is a data point, and $y \in \mathcal{Y} = \{1, \ldots, K\}$ is its corresponding label. We denote the clean training and test sets as $\mathcal{D}_c$ and $\mathcal{D}_t$, respectively. The goal of UEs is to transform $\mathcal{D}_c$ into an unlearnable dataset $\mathcal{D}_u$ by injecting perturbations $\boldsymbol{\delta}$ into each example, yielding $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}$. The perturbation $\boldsymbol{\delta}$ is constrained by $\|\boldsymbol{\delta}\|_p \leq \epsilon$, where $\|\cdot\|_p$

denotes the ℓ_p norm and ϵ is chosen to be sufficiently small to ensure imperceptibility. Consider a classification model $f = h \circ g$, where g is a fixed pretrained feature extractor and h_θ is a linear head with learnable parameters θ. Unlearnability is achieved when linear probing on $\mathcal{D}_u$ with frozen $g(\cdot)$ and trainable $h_\theta(\cdot)$ fails to generalize to $\mathcal{D}_t$, yielding clean test accuracy close to random guessing.

Framework Overview. Since the pretrained backbone is immune to perturbation manipulation during linear probing, we shift the design principle from feature disruption to boundary diversion, aiming to enforce discrepancies between decision boundaries learned on the perturbed training set and the clean test set. To achieve this, we design DIVERT as a two-step framework. In the first step (Section 3.2), we push samples into a semantically decoupled subspace to isolate them from the primary semantic features, thereby inducing spurious separability. In the second step (Section 3.3), we simulate the actual linear probing procedure to anticipate the perturbation-driven learning dynamics, which enables us to proactively divert the decision boundary onto the induced separability.

3.2 Inducing Spurious Separability in Semantic-Decoupled Subspace

Original Prototype. During linear probing, the pretrained feature extractor is able to encode images into cluster-wise embeddings (as illustrated in Figure 2), which naturally provides discriminative representations. To mislead models away from real prototypes, we construct synthetic class anchors in a semantically decoupled subspace. Specifically, we first calculate the D-dimensional class prototypes of the original data samples $\mathcal{D}_c$ through a pretrained feature extractor g:

$$\boldsymbol{c}_k = \frac{1}{N_k} \sum_{x \in \mathcal{D}_k} g(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{c}_k$ is the class prototype for class k, $\mathcal{D}_k = \{x_i\}_{i=1}^{N_k}$ is the set of training samples belonging to class $k \in \{1, 2, \ldots, K\}$, and N_k is the number of samples in that class.

We then define the D-dimensional global prototype $\boldsymbol{c}_g$ as:

$$\boldsymbol{c}_g = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{c}_k. \tag{2}$$

To capture the inter-class semantic structure while removing dataset-level bias, we construct the clean semantic basis matrix $\mathbf{M} \in \mathbb{R}^{K \times D}$ by centering each class prototype against the global prototype:

$$\mathbf{M}_k = \frac{\boldsymbol{c}_k - \boldsymbol{c}_g}{\|\boldsymbol{c}_k - \boldsymbol{c}_g\|_2}. \tag{3}$$

where $\mathbf{M}_k$ denotes the centered clean semantic basis of class k, $\|\cdot\|_2$ represents the ℓ_2 norm.

Semantic-decoupled Subspace Discovery. To induce spurious separability, we first identify a subspace orthogonal to the primary semantic directions and construct synthetic class anchors that are semantically decoupled from the genuine class prototypes. Concretely, we apply Singular Value Decomposition (SVD) to decompose the clean class semantic basis $\mathbf{M} \in \mathbb{R}^{K \times D}$. As shown in Eq. 4, this decomposition yields a set of orthogonal basis vectors that partition the feature space into a semantic subspace and its orthogonal complement, which provides an ideal foundation for constructing artificial target anchors.

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \tag{4}$$

where $\mathbf{U}$ and $\mathbf{\Sigma}$ denote the left singular vectors and the diagonal matrix of singular values, respectively. $\mathbf{V} \in \mathbb{R}^{D \times D}$ comprises the right singular vectors, forming an orthonormal basis for the feature space. Since the centered semantic basis matrix $\mathbf{M}$ has rank at most $K - 1$ (see Appendix A.3 for further justification), the front primary vectors $\mathbf{V}_{\text{sem}} := [v_1, \ldots, v_{K-1}] \in \mathbb{R}^{D \times (K-1)}$ capture the dominant inter-class semantic variation and constitute the semantic subspace, while the remaining complementary vectors $\mathbf{V}_{\text{comp}} := [v_K, \ldots, v_D] \in \mathbb{R}^{D \times (D-K+1)}$ is orthogonal to principal semantics and thus constitutes the semantic-decoupled subspace.

Attention-driven Synthetic Anchor Construction. To induce spurious linear separability away from true semantics, we construct synthetic class anchors within the semantic-decoupled subspace to serve as artificial targets, where we optimize perturbations to push unlearnable samples toward them. Concretely, we adopt a cross-attention mechanism [46, 47] to form K target anchors from all

complementary vectors within $\mathbf{V}_{\text{comp}}$. This design fully leverages the semantic-decoupled information and fosters dynamic interaction between complementary vectors, which adaptively explores the optimal synthetic anchors. Detailed formulation is described as follows:

$$T_k = \frac{\sum_{j=K}^{D} \alpha_{k,j}\, v_j}{\left\| \sum_{j=K}^{D} \alpha_{k,j}\, v_j \right\|_2}, \qquad \alpha_{k,j} = \frac{\exp\big(\tau \,\langle \text{query}_k, \text{key}_j \rangle\big)}{\sum_{j'=K}^{D} \exp\big(\tau \,\langle \text{query}_k, \text{key}_{j'} \rangle\big)}, \tag{5}$$

where for each class k, the target anchor T_k is formed as a learnable soft combination over the semantic-decoupled basis $\mathbf{V}_{\text{comp}}$, and the full anchor set is denoted by $\mathbf{T} = [T_1, \ldots, T_K] \in \mathbb{R}^{K \times D}$. query_k and key_j represent the ℓ_2-normalized query and key vectors obtained via learnable linear projections, and τ denotes the temperature coefficient. The basis vectors $\{v_j\}_{j=K}^{D}$ serve as fixed candidate vectors, while we update the attention module to dynamically refine the target anchors.

Anchor Diversity Constraint. During optimization, the attention mechanism may produce highly similar target anchors across different classes. Such overlap directs perturbed samples toward overlapping regions in the feature space, undermining the induced spurious inter-class separability. To prevent this potential anchor collapse and maintain distinct class representations, we introduce a structural diversity loss. This objective encourages mutual orthogonality among the synthetic anchors by applying a Gram-based penalty to their pairwise inner products[1]:

$$\min_{\mathbf{T}} \; \mathcal{L}_{\text{div}} = \sum_{i \neq j} \big(T_i T_j^{\top}\big)^2 = \big\| \mathbf{T}\mathbf{T}^{\top} - \mathbf{I}_K \big\|_F^2, \tag{6}$$

where T_i is the target anchor of class i, $\mathbf{I}_K$ denotes the identity matrix, and $\|\cdot\|_F$ represents the Frobenius norm. The diversity loss encourages distinct classes to be mapped to distinguishable directions in the semantic-decoupled subspace.

Perturbation-Anchor Co-adaptation. To push perturbed samples to the semantic-decoupled subspace, we design a shifting loss by calculating the cosine distance between perturbed samples and synthetic anchors in the feature space, as shown in Eq. 7. Meanwhile, to ease the shifting process given constrained perturbation budgets, we also treat the anchors as optimizable targets, where we formulate a min-min optimization framework with alternating updates, as shown in Eq. 8.

$$\mathcal{L}_{\text{shift}}(\boldsymbol{z}, y, \mathbf{T}; \boldsymbol{c}_g) = d\left(\boldsymbol{z} - \boldsymbol{c}_g, T_y\right), \tag{7}$$

$$\min_{\mathbf{T}} \left\{ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_c} \left[\min_{\|\boldsymbol{\delta}\|_\infty \leq \epsilon} \mathcal{L}_{\text{shift}}(\boldsymbol{z}, y, \mathbf{T}; \boldsymbol{c}_g) \right] + \lambda \mathcal{L}_{\text{div}} \right\}, \tag{8}$$

where $\boldsymbol{z} = g(\boldsymbol{x} + \boldsymbol{\delta})$ represents the encoded features of perturbed samples, $d(\mathbf{u}, \mathbf{v}) = 1 - \frac{\mathbf{u}^{\top}\mathbf{v}}{\|\mathbf{u}\|_2 \, \|\mathbf{v}\|_2}$ denotes the cosine distance. λ is the weight hyperparameter. The cosine distance is computed in the centered feature space, where the global prototype is subtracted to mitigate potential bias.

3.3 Diverting Decision Boundaries via Surrogate Linear Probing

Pushing perturbed samples toward synthetic anchors establishes spurious separability that deviates from the original class prototypes, laying the foundation for subsequent decision boundary diversion. To further refine perturbations, we anticipate the perturbation-driven learning dynamics and proactively simulate the real linear probing procedure. By involving a surrogate linear head in the optimization, we directly encourage perturbations to learn how to trick the model into learning the spurious separability for decision making. Concretely, we attach a surrogate linear head h_θ on top of the pretrained feature extractor g. Drawing on the error-minimization principle [7], we formulate a min-min optimization that alternately updates the perturbations and the linear head to reduce the classification loss, as shown in Eq. 9. Through this process, perturbations progressively divert the decision boundaries upon the established spurious separability basis, ultimately rendering distinct decision boundaries from genuine semantics and impeding the model's generalization to clean data.

$$\min_{\theta} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_c} \left[\min_{\|\boldsymbol{\delta}\|_\infty \leq \epsilon} \mathcal{L}_{\text{CE}}(h_\theta(\boldsymbol{z}), y) \right], \tag{9}$$

where $\boldsymbol{z} = g(\boldsymbol{x} + \boldsymbol{\delta})$ represents the encoded features of perturbed samples, $g(\cdot)$, $h_\theta(\cdot)$ denotes the feature extractor and linear head of the surrogate, and $\mathcal{L}_{\text{CE}}$ indicates the cross-entropy loss.

[1] For brevity, we write $\mathbf{T}$ as the optimization variable in Eq. 6. In practice, the gradient is back-propagated to the underlying attention parameters that compose $\mathbf{T}$.

3.4 Overall Optimization Procedure

The complete optimization is formulated as a progressive two-step framework, as shown in Eq. 10. In the first step, we attempt to shift perturbed samples into the semantic-decoupled subspace by alternately updating the perturbations $\boldsymbol{\delta}$ and the synthetic anchors $\mathbf{T}$ to minimize the shifting loss $\mathcal{L}_{\text{shift}}$. This step establishes the geometric basis to alter the decision boundary. In the second step, we incorporate a surrogate linear head h_θ and jointly optimize the perturbations $\boldsymbol{\delta}$ alongside both h_θ and $\mathbf{T}$, aiming to actively divert the learned decision boundary onto the spurious separability built in the first step. Through this two-step procedure, our method is able to generate UEs that enforce boundary discrepancies and remain effective against linear probing.

$$\begin{cases} \text{S1}: \min\limits_{\mathbf{T}} \left\{ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_c} \left[\min\limits_{\|\boldsymbol{\delta}\|_\infty \le \epsilon} \mathcal{L}_{\text{shift}}(\boldsymbol{z}, y, \mathbf{T}; \boldsymbol{c}_g) \right] + \lambda \mathcal{L}_{\text{div}} \right\}, \\ \text{S2}: \min\limits_{\theta,\mathbf{T}} \left\{ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_c} \left[\min\limits_{\|\boldsymbol{\delta}\|_\infty \le \epsilon} \Big(\mathcal{L}_{\text{shift}}\big(\boldsymbol{z}, y, \mathbf{T}; \boldsymbol{c}_g\big) + \mathcal{L}_{\text{CE}}\big(h_\theta(\boldsymbol{z}), y\big) \Big) \right] + \lambda \mathcal{L}_{\text{div}} \right\}, \end{cases} \tag{10}$$

where $\boldsymbol{z} = g(\boldsymbol{x} + \boldsymbol{\delta})$ denotes the perturbed features and λ is the weight hyperparameter.

4 Experiment

In this section, we evaluate the proposed method under linear probing. We then examine its transferability across diverse pretraining regimes, datasets, and network architectures, and assess its resistance against defenses and applicability under the conventional fully trainable setting. Finally, we present ablation studies along with qualitative visualization analyses.

4.1 Experimental Setup

Datasets and Models. We employ four standard benchmark datasets: CIFAR-10 [48], CIFAR-100 [48], SVHN [49], and Tiny-ImageNet [50]. We evaluate a variety of network architectures, including ResNet-18 [51], ResNet-50 [51], VGG-11 [52], DenseNet-121 [53], GoogLeNet [54], and ViT [55]. Unless otherwise specified, all evaluations are conducted via linear probing with an ImageNet-pretrained backbone, where the pretrained feature extractor is frozen and only the linear head is trainable. The linear head is reinitialized to match the number of classes in each downstream dataset. We measure unlearnability by the classification accuracy on the clean test set, where lower accuracy typically indicates stronger protection.

Baseline Methods. We compare DIVERT with various methods, including error-minimizing noises (EMN) [7], transferable unlearnable examples (TUE) [32], robust error-minimizing noises (REM) [56], linear separable perturbations (LSP) [57], convolution-based unlearnable datasets (CUDA) [58], game unlearnable example (GUE) [59], and versatile unlearnable generator (VTG) [60].

Implementation Details. We generate perturbations under an ℓ_∞ budget of $8/255$, in accordance with prior works [7, 32, 59]. During optimization, we employ an ImageNet-pretrained backbone with a reinitialized linear head as the surrogate model, trained using the SGD optimizer with a learning rate of 0.001. The training epoch is set to 60, where we craft perturbations by conducting PGD attack [61] with a step size of $0.8/255$ in a class-wise manner. The balancing hyperparameter λ is set to 0.5 by default. Our method is implemented with PyTorch and runs on a single NVIDIA L40S GPU. All results are reported as averages over five runs with different random seeds.

4.2 Main Results

Unlearnability on Various Datasets. To demonstrate the efficacy of DIVERT, we evaluate our method against a diverse set of ImageNet-pretrained backbones. For each backbone, perturbations are optimized and evaluated under the same architecture to ensure a fair comparison. As reported in Table 1, our method consistently outperforms all baseline UE methods across both CNN-based and ViT-based backbones on CIFAR-10, CIFAR-100, SVHN, and Tiny-ImageNet. In most cases, our method effectively reduces test accuracy to random-guessing levels. These results highlight the superior strength of DIVERT in generating perturbations that push perturbed samples to the semantic-decoupled subspace and shift decision boundaries, successfully degrading the generalization of models to clean test data and protecting user data.

Table 1: Test accuracy (%) ↓ comparison across different backbones against linear probing.

Dataset	Backbone	Clean	EMN	TUE	REM	LSP	CUDA	GUE	VTG	**Ours**
CIFAR-10	ResNet-18	80.73	44.54	71.80	51.81	66.89	60.43	80.39	70.82	**13.21**
	ResNet-50	85.16	22.89	67.17	23.64	75.33	64.77	84.78	78.95	**14.29**
	DenseNet-121	81.94	48.61	56.78	30.30	70.02	58.52	81.72	77.06	**11.21**
	VGG-11	72.01	62.73	62.15	41.43	70.55	63.99	71.82	64.94	**11.63**
	ViT	85.01	70.55	83.70	40.57	83.35	65.47	84.40	74.66	**22.42**
CIFAR-100	ResNet-18	58.12	18.86	44.37	28.66	54.64	44.48	57.05	40.77	**3.21**
	ResNet-50	63.23	29.48	29.02	40.01	50.64	48.35	62.42	56.65	**4.13**
	DenseNet-121	58.66	26.43	19.28	20.20	39.14	39.05	57.58	39.61	**4.41**
	VGG-11	44.82	31.07	31.91	22.35	42.13	38.56	44.45	43.73	**1.82**
	ViT	57.86	39.74	57.50	33.46	48.05	39.22	55.56	56.15	**12.93**
SVHN	ResNet-18	47.97	10.58	16.68	21.42	24.50	20.66	42.47	19.38	**9.16**
	ResNet-50	54.04	11.09	17.71	48.01	21.22	19.79	50.90	30.07	**9.56**
	DenseNet-121	48.85	19.85	16.91	9.19	20.16	19.08	44.84	25.84	**6.94**
	VGG-11	45.02	30.34	26.76	25.29	40.16	20.03	42.91	25.87	**9.58**
	ViT	51.07	45.90	50.96	17.69	33.51	22.00	49.13	19.02	**15.53**
Tiny-ImageNet	ResNet-18	57.70	25.69	27.09	42.81	54.12	34.20	55.97	57.21	**7.09**
	ResNet-50	63.66	16.33	53.39	65.77	60.87	39.14	63.49	35.72	**4.74**
	DenseNet-121	60.57	40.69	21.81	43.05	57.45	37.08	60.24	55.60	**7.20**
	VGG-11	47.96	42.49	43.09	40.10	47.51	31.12	47.91	47.84	**11.73**
	ViT	68.19	62.90	62.77	60.56	67.17	51.12	67.97	68.08	**10.02**

Table 2: Transferability across diverse pretraining priors, where perturbations are generated with an ImageNet-pretrained surrogate ResNet-50 and applied to different pretrained backbones. Results are obtained on CIFAR-10.

Framework	Pretraining Data	Paradigm	Clean	TUE	REM	LSP	GUE	VTG	**Ours**
Standard	Places365	Supervised	78.08	64.54	45.26	62.13	38.37	20.14	**19.43**
MoCo v3	ImageNet	Self-Supervised	87.88	66.16	15.58	64.13	36.26	15.28	**12.18**
CLIP	WIT-400M	Self-Supervised	61.72	22.52	14.49	36.90	23.28	43.09	**10.02**

Transferability Across Diverse Pretraining Regimes. We optimize perturbations by employing surrogates pretrained on ImageNet [1] under supervised learning. However, the pretrained backbone adopted by the unauthorized user for linear probing is unknown and may differ in either the pretraining data or the training paradigm. To assess the transferability of our method under such variability, we conduct evaluations across three ResNet-50 backbones pretrained under three distinct regimes: standard supervised pretraining on Places365 [62], self-supervised pretraining on ImageNet (MoCo v3 [63]), and image-text contrastive learning on WIT-400M (CLIP [14]). As reported in Table 2, our method consistently reduces test accuracy across all three regimes, demonstrating that DIVERT successfully diverts the decision boundary regardless of the underlying pretraining representations.

Cross-Dataset Transferability. In practice, the data distribution encountered by unauthorized users may differ from that used for perturbation generation [64, 65, 66]. Therefore, we further evaluate the transferability of our method across different datasets, where we optimize perturbations on Tiny-ImageNet and evaluate them on CIFAR-10, CIFAR-100, and SVHN. Since our method operates in a class-level manner, we randomly select perturbations matching the number of classes in the target dataset. For sample-wise UE methods (e.g., TUE, REM), we select perturbations according to the number of images in the target dataset. For generator-based methods (GUE and VTG), we directly apply the trained generator to produce perturbations for the target data. As presented in Table 3, our method outperforms all baseline methods in rendering the target dataset unlearnable, demonstrating superior cross-dataset transferability.

Cross-Architecture Transferability. Following previous studies [7, 32, 59], we employ a surrogate model to assist the perturbation optimization procedure. In practice, however, the network architecture that the unauthorized user adopts for training is unknown, requiring perturbations to transfer across different network architectures. To assess this transferability, we optimize perturbations using a surrogate ResNet-18 and evaluate them across diverse pretrained backbones spanning both CNN and ViT under linear probing. As shown in Table 4, we observe that DIVERT consistently reduces

Table 3: Test accuracy (%) ↓ of cross-dataset transferability, where perturbations are generated on Tiny-ImageNet with ResNet-18 and applied to different datasets.

Source	Target	Clean	EMN	TUE	REM	LSP	GUE	VTG	**Ours**
Tiny-ImageNet	CIFAR-10	80.73	35.39	57.60	74.57	65.68	77.19	80.67	**11.75**
	CIFAR-100	58.12	16.89	21.41	45.30	40.02	55.01	57.55	**4.24**
	SVHN	47.97	14.14	26.51	37.02	23.81	40.63	47.34	**9.55**

Table 4: Test accuracy (%) ↓ of cross-architecture transferability on CIFAR-10 with ResNet-18.

Method	Network Architecture				
	ResNet-50	VGG-11	DenseNet-121	GoogLeNet	ViT
Clean	85.16	72.01	81.94	81.54	85.01
EMN	27.65	25.34	22.62	24.29	66.18
TUE	63.59	68.30	67.53	67.20	81.68
REM	24.32	37.32	42.99	46.79	72.78
GUE	85.24	72.18	81.26	81.23	84.86
VTG	75.10	70.94	74.12	76.51	84.58
Ours	**23.73**	**20.47**	**18.70**	**14.05**	**22.63**

test accuracy to near chance level across all tested architectures. This consistent effectiveness across diverse backbones demonstrates that our method does not rely on architecture-specific feature representations. Rather, the induced spurious separability and decision boundary diversion generalize broadly, validating the practical applicability of DIVERT.

Resilience Against Defenses. In practice, unauthorized parties may apply defensive strategies to circumvent unlearnability. We therefore evaluate the robustness of UEs to common defenses, including widely-used data augmentations (Cutout [67], CutMix [68], and Mixup [69]) and JPEG [70] compression with a quality factor of 50 (see Appendix C.6 for full results across quality factors). As shown in Table 5, even under additional defenses, our method still maintains strong unlearnability, consistently degrading test accuracy and demonstrating superior robustness compared to baselines.

Table 5: Test accuracy (%) ↓ under additional defenses against linear probing with ResNet-18, where CIFAR-10 is utilized as the target dataset.

Method	w/o	Cutout	CutMix	Mixup	JPEG
Clean	80.73	79.50	77.76	77.80	62.08
EMN	44.54	43.21	40.30	46.43	32.32
TUE	71.80	66.40	65.12	64.86	39.05
REM	51.81	53.67	54.40	52.37	47.21
LSP	66.89	66.17	67.38	68.56	32.17
VTG	70.82	78.55	75.87	75.76	49.72
Ours	**13.21**	**17.20**	**17.67**	**19.07**	**23.15**

4.3 Further Analyses

Ablation Studies. To examine the impact of each component within the proposed boundary diversion framework, we conduct detailed ablation studies. As reported in Table 6, $\mathcal{L}_{\text{shift}}$ provides the foundation by pushing perturbed features into the semantically orthogonal subspace, yielding initial unlearnability. Building upon this, $\mathcal{L}_{\text{div}}$ prevents anchor collapse and promotes diversity among the synthetic class anchors within the semantic-decoupled subspace, further enhancing unlearnability. Finally, $\mathcal{L}_{\text{CE}}$ explicitly drives perturbations to divert the decision boundary, achieving the strongest data protection.

Table 6: Ablation study of DIVERT on ResNet-18.

$\mathcal{L}_{\text{shift}}$	$\mathcal{L}_{\text{div}}$	$\mathcal{L}_{\text{CE}}$	CIFAR-10	CIFAR-100	SVHN
✓			23.92	7.05	15.45
✓	✓		18.86	5.27	11.37
✓	✓	✓	**13.21**	**3.21**	**9.16**

Effectiveness Under the Standard Fully Trainable Setting. The core idea of our method is to push perturbed samples toward the semantic-decoupled subspace and divert the decision boundary, which naturally extends to the conventional UE setting with a fully trainable target model. As reported in Table 7, our method exhibits remarkable efficacy and achieves the lowest test accuracy among all

Table 7: Test accuracy (%) ↓ under the standard fully trainable evaluation setting on CIFAR-10.

Method	ResNet-18	ResNet-50	VGG-11	DenseNet-121	ViT
EMN	16.42	13.45	16.93	14.71	20.79
TUE	10.67	12.23	12.79	17.60	23.25
REM	15.18	25.57	47.68	41.54	57.21
LSP	13.54	15.94	17.15	17.47	23.19
GUE	13.25	12.97	23.89	13.71	23.10
VTG	9.99	10.78	10.45	10.16	12.73
Ours	**9.76**	**10.02**	**9.67**	**9.98**	**11.11**

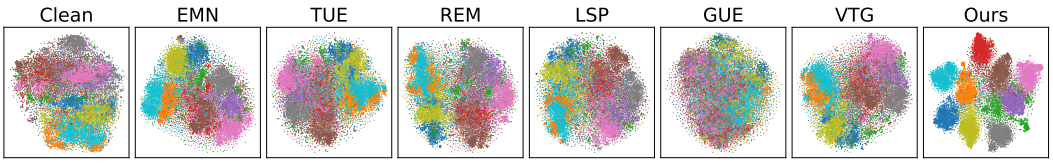


Figure 2: t-SNE visualization of the perturbed training samples on CIFAR-10.

baselines, demonstrating strong applicability under both the standard fully trainable setting and the newly proposed linear probing setting.

4.4 Qualitative Evaluation

Visualization of Perturbed Features. To visually demonstrate the effect of our method, we visualize the features of perturbed training samples extracted by an ImageNet-pretrained model via t-SNE [71]. As illustrated in Figure 2, baseline UE methods yield feature distributions closely resembling those of clean data, indicating their perturbations fail to alter the underlying class structure. In contrast, our method produces markedly distinct clusters that deviate from the original distribution, confirming that pushing perturbed samples into the orthogonal semantic-decoupled subspace effectively reshapes the feature geometry and lays the foundation for diverting the decision boundary.

Visualization of Unlearnable Samples. Besides effective unlearnability, perturbations also have to be small enough to remain invisible to human eyes. In accordance with previous studies [7, 32], we have explicitly constrained the budget of perturbations to be $8/255$. To provide a more vivid qualitative assessment, we visualize our perturbed images on CIFAR-10 and make comparisons with existing UE methods. As illustrated in Figure 3, we observe that the perturbations optimized by our method are generally imperceptible to humans. We further present more visualizations of the perturbed images on CIFAR-100, SVHN, and Tiny-ImageNet for a more comprehensive analysis, which is shown in Appendix F.

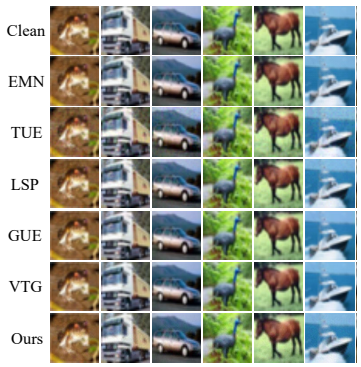


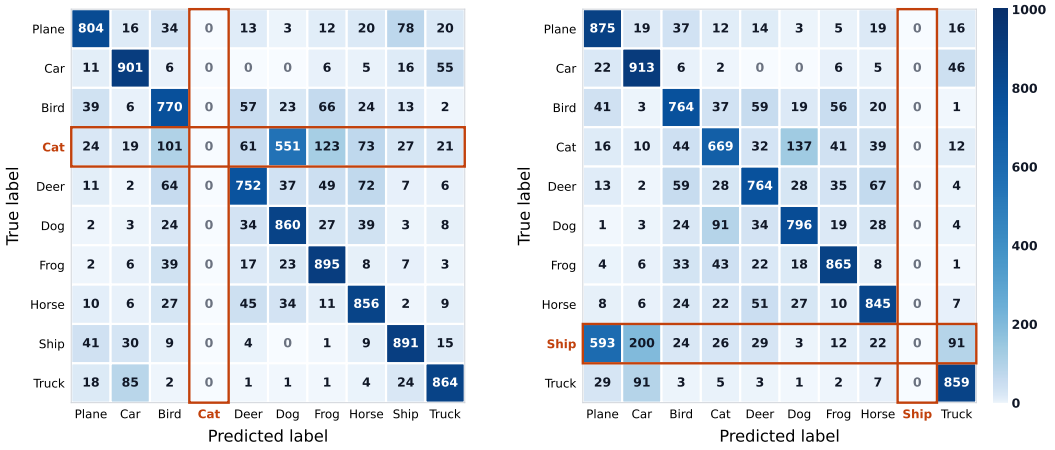
Figure 3: Unlearnable samples on CIFAR-10.

4.5 Partial Protection

Partial Dataset Protection. Standard UE evaluations typically assume that the entire training set is protected. However, real-world applications often involve partially poisoned datasets. To address this practical constraint, we follow previous studies [7, 58] to systematically evaluate the unlearnability of various methods under different protection ratios, with results summarized in Table 8. We observe that combining perturbed data with clean samples does not yield a substantial increase in test accuracy relative to the clean-only baseline. This marginal performance difference indicates that the models are unable to extract additional semantic knowledge from the protected subset, confirming that unlearnability is effectively maintained even in mixed training environments. Furthermore, our

Table 8: Test accuracy (%) ↓ comparison with different poisoning ratios against linear probing, where we utilize ResNet-18 on CIFAR-10 during evaluation.

Poison Ratio r	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Clean $(1-r)$	80.62	80.46	79.82	80.32	80.18	79.83	79.53	78.73	77.66	–
EMN	80.54	80.34	80.16	81.52	80.58	79.98	79.16	77.37	74.45	44.54
TUE	80.52	80.27	80.13	79.84	79.31	78.57	77.22	75.76	73.96	71.80
REM	80.65	80.36	80.39	80.20	80.16	79.89	79.87	78.44	75.48	51.81
LSP	80.74	80.49	80.47	80.11	79.86	79.79	79.32	78.37	77.22	66.89
VTG	80.83	80.71	80.50	80.48	80.25	80.21	80.14	80.53	79.30	70.82
Ours	**80.39**	**80.22**	**79.76**	**79.42**	**79.20**	**78.34**	**77.07**	**75.72**	**71.01**	**13.21**



(a) Protecting "cat" only **(b) Protecting "ship" only**

Figure 4: Confusion matrices for partial class protection on CIFAR-10, where perturbations are applied only to training samples from the "Cat" or "Ship" class.

proposed method consistently provides the most rigorous protection across all tested ratios. It strictly constrains the test accuracy below the clean baseline, demonstrating superior effectiveness.

Partial Class Protection. We further evaluate DIVERT on CIFAR-10 by perturbing all training samples from either the "Cat" or "Ship" class while leaving the remaining classes clean. As shown in Figure 4, clean-test accuracy on the protected class drops to 0% in both settings, with all test samples from that class misclassified. These results demonstrate that DIVERT can provide effective class-specific protection even when full-dataset protection is difficult to achieve.

5 Conclusion

In this paper, we present the first study of UEs under linear probing and uncover a fundamental vulnerability. We empirically show that previous UEs operate by disrupting clean feature representations, a mechanism that struggles when the backbone is frozen and leaves representations immune to such disruption. To address this limitation, we introduce DIVERT, a two-step optimization framework. DIVERT first constructs synthetic class anchors within the semantic-decoupled subspace and pushes samples toward them to induce spurious separability. It then refines the perturbations under a simulated linear probing procedure to actively drive decision boundary diversion. Extensive experiments demonstrate that DIVERT achieves strong unlearnability across diverse datasets and backbones.

Limitations. Our work explores unlearnability within the image modality. Since linear probing can also be applied in other modalities and potentially affect unlearnability, extending DIVERT accordingly is left for future work. Moreover, partial protection remains a common challenge for UE methods, as the remaining clean samples can sustain high overall accuracy. Nevertheless, DIVERT-protected samples provide only little additional learnable semantic information.

Acknowledgments and Disclosure of Funding

This work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), Discovery Grants program.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[3] Hao Zheng, Zhigang Hu, Liu Yang, Aikun Xu, Meiguang Zheng, Ce Zhang, and Keqin Li. Multifeature collaborative fusion network with deep supervision for sar ship classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

[4] Hang Zhou, Jiale Cai, Yuteng Ye, Yonghui Feng, Chenxing Gao, Junqing Yu, Zikai Song, and Wei Yang. Video anomaly detection with motion and appearance guided patch diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10761–10769, 2025.

[5] Ruiyi Fang, Jingyu Zhao, Shuo Wang, Ruizhi Pu, Bingheng Li, Jiale Cai, Zhihao Li, Zihao Jing, Jian Zhu, Song Tang, et al. Saga: Structural aggregation guided alignment with dynamic view and neighborhood order selection for multiview graph domain adaptation. In *International Conference on Learning Representations*, volume 2026, pages 95585–95607, 2026.

[6] Hongsheng Zhang, Zhong Ji, Di Wang, and Yanwei Pang. Interleaved prompt generation for continual instruction tuning of multimodal large language model. *IEEE Transactions on Image Processing*, 2026.

[7] 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.

[8] Pedro Sandoval-Segura, Vasu Singla, Jonas Geiping, Micah Goldblum, and Tom Goldstein. What can we learn from unlearnable datasets? *Advances in Neural Information Processing Systems*, 36:75372–75391, 2023.

[9] Yifan Zhu, Yibo Miao, Yinpeng Dong, and Xiao-Shan Gao. Toward availability attacks in 3d point clouds. In *Forty-first International Conference on Machine Learning*, 2024.

[10] Minghui Li, Xianlong Wang, Zhifei Yu, Shengshan Hu, Ziqi Zhou, Longling Zhang, and Leo Yu Zhang. Detecting and corrupting convolution-based unlearnable examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18403–18411, 2025.

[11] Jinlin Liu, Wei Chen, and Xiaojin Zhang. Perturbation-induced linearization: Constructing unlearnable data with solely linear classifiers. In *The Fourteenth International Conference on Learning Representations*, 2026.

[12] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[14] 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*, pages 8748–8763. PMLR, 2021.

[15] Ruizhi Pu, Lixing Yu, Shaojie Zhan, Gezheng Xu, Fan Zhou, Charles X Ling, and Boyu Wang. Fedelr: When federated learning meets learning with noisy labels. *Neural Networks*, 187:107275, 2025.

[16] Xiuting Weng, Ruizhi Pu, Yuanhang Yao, Kun Yue, Zhiwen Tang, and Lixing Yu. Gdfa: Geometry-driven federated unlearning with directional task vector alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10346–10356, 2026.

[17] Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. Badchain: Backdoor chain-of-thought prompting for large language models. In *International Conference on Learning Representations*, volume 2024, pages 28289–28316, 2024.

[18] Changjian Shui, Gezheng Xu, Qi Chen, Jiaqi Li, Charles X Ling, Tal Arbel, Boyu Wang, and Christian Gagné. On learning fairness and accuracy on multiple subgroups. *Advances in Neural Information Processing Systems*, 35:34121–34135, 2022.

[19] Ruizhi Pu, Gezheng Xu, Ruiyi Fang, Bing-Kun Bao, Charles Ling, and Boyu Wang. Leveraging group classification with descending soft labeling for deep imbalanced regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19978–19985, 2025.

[20] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.

[21] Junfeng Guo, Yiming Li, Lixu Wang, Shu-Tao Xia, Heng Huang, Cong Liu, and Bo Li. Domain watermark: Effective and harmless dataset copyright protection is closed at hand. *Advances in Neural Information Processing Systems*, 36:54421–54450, 2023.

[22] Yuhao Sun, Lingyun Yu, Hongtao Xie, Jiaming Li, and Yongdong Zhang. Diffam: Diffusion-based adversarial makeup transfer for facial privacy protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24584–24594, 2024.

[23] Ziming Hong, Zhenyi Wang, Li Shen, Yu Yao, Zhuo Huang, Shiming Chen, Mingming Gong, Tongliang Liu, et al. Improving non-transferable representation learning by harnessing content and style. In *International Conference on Learning Representations*, volume 2024, pages 27472–27499, 2024.

[24] Yuxin Huang, Ziming Hong, Mingming Gong, Wanyu Wang, Jing Zhang, and Tongliang Liu. Delving into the temporal challenges of unified video protection against image-to-video and fine-tuning-based customization. *arXiv preprint arXiv:2607.13336*, 2026.

[25] Jiahao Li, Yiqiang Chen, Yunbing Xing, Yang Gu, and Xiangyuan Lan. K-space bispectrum steganography for robust unlearnable data. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11492–11501, 2025.

[26] Zhihao Li, Gezheng Xu, Jiale Cai, Ruiyi Fang, Di Wu, Qicheng Lao, Charles Ling, and Boyu Wang. When priors backfire: On the vulnerability of unlearnable examples to pretraining. In *International Conference on Learning Representations*, pages 17454–17474, 2026.

[27] Yifan Zhu, Yibo Miao, Yinpeng Dong, and Xiao-Shan Gao. Why do unlearnable examples work: A novel perspective of mutual information. In *The Fourteenth International Conference on Learning Representations*, 2026.

[28] Jiale Cai, Gezheng Xu, Zhihao Li, Ruiyi Fang, Ruizhi Pu, Di Wu, Qicheng Lao, Charles Ling, and Boyu Wang. FUSE: Full-spectrum unlearnable examples via spectral equalization. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, pages 10764–10788. PMLR, 2026.

[29] Xianlong Wang, Hangtao Zhang, Wenbo Pan, Ziqi Zhou, Changsong Jiang, Li Zeng, and Xiaohua Jia. Dual-branch robust unlearnable examples. In *Forty-third International Conference on Machine Learning*, 2026.

[30] Chengshuai Zhao, Zhen Tan, Dawei Li, Zhiyuan Yu, and Huan Liu. To see is not to learn: Protecting multi-modal data from unauthorized fine-tuning of large vision-language model. *arXiv preprint arXiv:2605.14291*, 2026.

[31] Chengshuai Zhao, Pingchuan Ma, Dawei Li, Bohan Jiang, Zhiyuan Yu, Zhen Tan, and Huan Liu. Textcloak: Thwarting unauthorized llm exploitation via rl-driven unlearnable text. *arXiv preprint arXiv:2607.28862*, 2026.

[32] Jie Ren, Han Xu, Yuxuan Wan, Xingjun Ma, Lichao Sun, and Jiliang Tang. Transferable unlearnable examples. In *The Eleventh International Conference on Learning Representations*, 2023.

[33] Pedro Sandoval-Segura, Vasu Singla, Jonas Geiping, Micah Goldblum, Tom Goldstein, and David Jacobs. Autoregressive perturbations for data poisoning. *Advances in Neural Information Processing Systems*, 35:27374–27386, 2022.

[34] Shutong Wu, Sizhe Chen, Cihang Xie, and Xiaolin Huang. One-pixel shortcut: On the learning preference of deep neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.

[35] Derui Wang, Minhui Xue, Bo Li, Seyit Camtepe, and Liming Zhu. Provably unlearnable data examples. In *The Network and Distributed System Security (NDSS) Symposium*, 2025.

[36] Hao He, Kaiwen Zha, and Dina Katabi. Indiscriminate poisoning attacks on unsupervised contrastive learning. In *International Conference on Learning Representations*, 2023.

[37] Yihan Wang, Yifan Zhu, and Xiao-Shan Gao. Efficient availability attacks against supervised and contrastive learning simultaneously. *Advances in Neural Information Processing Systems*, 37:72872–72900, 2024.

[38] 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 Conference on Computer Vision and Pattern Recognition*, pages 3984–3993, 2023.

[39] Chaochao Chen, Jiaming Zhang, Yuyuan Li, and Zhongxuan Han. One for all: A universal generator for concept unlearnability via multi-modal alignment. In *International Conference on Machine Learning*, pages 7700 – 7711, 2024.

[40] Ye Sun, Hao Zhang, Tiehua Zhang, Xingjun Ma, and Yu-Gang Jiang. Unseg: One universal unlearnable example generator is enough against all image segmentation. *Advances in Neural Information Processing Systems*, 37:79168–79193, 2024.

[41] Xianlong Wang, Minghui Li, Wei Liu, Hangtao Zhang, Shengshan Hu, Yechao Zhang, Ziqi Zhou, and Hai Jin. Unlearnable 3d point clouds: Class-wise transformation is all you need. *Advances in Neural Information Processing Systems*, 37:99404–99432, 2024.

[42] Chenyu Yi, Ruohan Meng, Haohang Peng, Bingquan Shen, and Alex C Kot. Towards effective and robust unlearnable examples against object detection. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 91–96. IEEE, 2025.

[43] Bo Li, Kun Zhang, Xi Chen, and Richang Hong. Syntax-oriented shortcut: A syntax level perturbing algorithm for preventing text data from being learned. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[44] Yunjie Ge, Ruxi Gu, Yuchen Liu, Lingchen Zhao, Bo Du, and Qian Wang. When unlearnable examples cooperate with watermarking: A dual voice data protection against unauthorized exploitation. *IEEE Transactions on Dependable and Secure Computing*, 2025.

[45] Lubin Meng, Xue Jiang, Jian Huang, Wei Li, Hanbin Luo, and Dongrui Wu. User identity protection in eeg-based brain–computer interfaces. *IEEE transactions on neural systems and rehabilitation engineering*, 31:3576–3586, 2023.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[47] Qiuhao Zeng, Jierui Huang, Peng Lu, Gezheng Xu, Boxing Chen, Charles Ling, and Boyu Wang. Zeta: Leveraging z-order curves for efficient top-k attention. In *International Conference on Learning Representations*, volume 2025, pages 72733–72757, 2025.

[48] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[49] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.

[50] Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[53] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[54] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[55] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *European conference on computer vision*, pages 68–85. Springer, 2022.

[56] Shaopeng Fu, Fengxiang He, Yang Liu, Li Shen, and Dacheng Tao. Robust unlearnable examples: Protecting data against adversarial learning. In *International Conference on Learning Representations*, 2022.

[57] Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Availability attacks create shortcuts. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2367–2376, 2022.

[58] Vinu Sankar Sadasivan, Mahdi Soltanolkotabi, and Soheil Feizi. Cuda: Convolution-based unlearnable datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3862–3871, 2023.

[59] Shuang Liu, Yihan Wang, and Xiao-Shan Gao. Game-theoretic unlearnable example generator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21349–21358, 2024.

[60] Zhihao Li, Jiale Cai, Gezheng Xu, Hao Zheng, Qiuyue Li, Fan Zhou, Shichun Yang, Charles Ling, and Boyu Wang. Versatile transferable unlearnable example generator. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[61] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[62] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[63] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.

[64] Qiuhao Zeng, Wei Wang, Fan Zhou, Charles Ling, and Boyu Wang. Foresee what you will learn: Data augmentation for domain generalization in non-stationary environment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11147–11155, 2023.

[65] Qiuhao Zeng, Changjian Shui, Long-Kai Huang, Peng Liu, Xi Chen, Charles Ling, and Boyu Wang. Latent trajectory learning for limited timestamps under distribution shift over time. In *The Twelfth International Conference on Learning Representations*, 2024.

[66] Gezheng Xu, Qi CHEN, QIUHAO Zeng, Charles Ling, and Boyu Wang. Discretized density-guided source-free domain adaptation for regression. In *Forty-third International Conference on Machine Learning*, 2026.

[67] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[68] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032, 2019.

[69] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[70] Zhuoran Liu, Zhengyu Zhao, and Martha Larson. Image shortcut squeezing: Countering perturbative availability poisons with compression. In *International Conference on Machine Learning*, pages 22473–22487. PMLR, 2023.

[71] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[72] 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*, pages 554–561, 2013.

[73] Ruiyi Fang, Bingheng Li, Qiuhao Zeng, Nima Hosseini Dashtbayaz, Ruizhi Pu, Charles Ling, Boyu Wang, et al. On the benefits of attribute-driven graph domain adaptation. In *International Conference on Learning Representations*, volume 2025, pages 30762–30782, 2025.

[74] Ruiyi Fang, Bingheng Li, Jingyu Zhao, Ruizhi Pu, QIUHAO Zeng, Gezheng Xu, Charles Ling, and Boyu Wang. Homophily enhanced graph domain adaptation. In *Forty-second International Conference on Machine Learning*, 2025.

[75] Gezheng Xu, Li Yi, Pengcheng Xu, Jiaqi Li, Ruizhi Pu, Changjian Shui, A Ian McLeod, Boyu Wang, and Charles Ling. Unraveling the mysteries of label noise in source-free domain adaptation: Theory and practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):4007–4023, 2025.

[76] Yan Zhang, Zhong Ji, Di Wang, Yanwei Pang, and Xuelong Li. User: Unified semantic enhancement with momentum contrast for image-text retrieval. *IEEE Transactions on Image Processing*, 33:595–609, 2024.

[77] Zhong Ji, Zhihao Li, Yan Zhang, Haoran Wang, Yanwei Pang, and Xuelong Li. Hierarchical matching and reasoning for multi-query image retrieval. *Neural Networks*, 173:106200, 2024.

[78] YuTeng Ye, Jiale Cai, Hang Zhou, Guanwen Li, Youjia Zhang, Zikai Song, Chenxing Gao, Junqing Yu, and Wei Yang. Progressive text-to-image diffusion with soft latent direction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6693–6701, 2024.

[79] Zhong Ji, Zhihao Li, Yan Zhang, Yanwei Pang, and Xuelong Li. Visual semantic contextualization network for multi-query image retrieval. *IEEE Transactions on Multimedia*, 27:7067–7080, 2025.

[80] Hui Guo, Boyu Wang, and Grace Yi. Label correction of crowdsourced noisy annotations with an instance-dependent noise transition model. *Advances in Neural Information Processing Systems*, 36:347–386, 2023.

[81] Hui Guo, Grace Y Yi, and Boyu Wang. Learning from noisy labels via conditional distributionally robust optimization. *Advances in Neural Information Processing Systems*, 37:82627–82672, 2024.

[82] Hui Guo, Grace Y Yi, and Boyu Wang. Addressing both variable selection and misclassified responses with parametric and semiparametric methods. *Bernoulli*, 32(2):1303–1327, 2026.

[83] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 20617–20642, 2024.

[84] Marco Fumero, Marco Pegoraro, Valentino Maiorca, Francesco Locatello, and Emanuele Rodolà. Latent functional maps: a spectral framework for representation alignment. *Advances in Neural Information Processing Systems*, 37:66178–66203, 2024.

[85] Zhihao Li, Yao Du, Yang Liu, Yan Zhang, Yufang Liu, Mengdi Zhang, Xunliang Cai, Charles Ling, and Boyu Wang. Eagle: Elevating geometric reasoning through llm-empowered visual instruction tuning. *arXiv preprint arXiv:2408.11397*, 2024.

[86] Yan Zhang, Zhong Ji, Yanwei Pang, Jungong Han, and Xuelong Li. Modality-experts coordinated adaptation for large multimodal models. *Science China Information Sciences*, 67(12):220107, 2024.

[87] Yan Zhang, Zhijin Qin, Feng Xu, Guiguang Ding, and Jungong Han. CofactGVR: Counterfactual intervention for grounded visual reasoning. In *Forty-third International Conference on Machine Learning*, 2026.

Appendix: Table of Contents

A More Details of the Proposed Method 17

A.1 More Qualitative Evidence of the Inherent Challenge against Linear Probing 17

A.2 Details for Quantifying Semantic Shift 17

A.3 Justification for the Rank of the Semantic Basis Matrix 18

A.4 Choice of SVD for Semantic Decomposition 18

A.5 Empirical Evidence of Boundary Diversion 18

A.6 Comparison with LSP 19

B More Experimental Details 19

B.1 Datasets 19

B.2 Implementation Details 19

C More Experiments 20

C.1 Evaluation on Fine-Grained Datasets 20

C.2 Protection without Ground-Truth Labels 20

C.3 Surrogate Model Sensitivity 20

C.4 Hyperparameter Sensitivity of Weight Coefficient 21

C.5 Impact of Perturbation Budget 21

C.6 Complete JPEG Defense 21

D More Analyses 22

D.1 Performance Gap between CNN and ViT 22

D.2 Mechanism of Cross-Architecture Transferability 22

E Computational Cost 23

F More Visualizations of Perturbed Images 24

G Discussion on Potential Implications 24

H Potential Societal Impact 25

A More Details of the Proposed Method

A.1 More Qualitative Evidence of the Inherent Challenge against Linear Probing

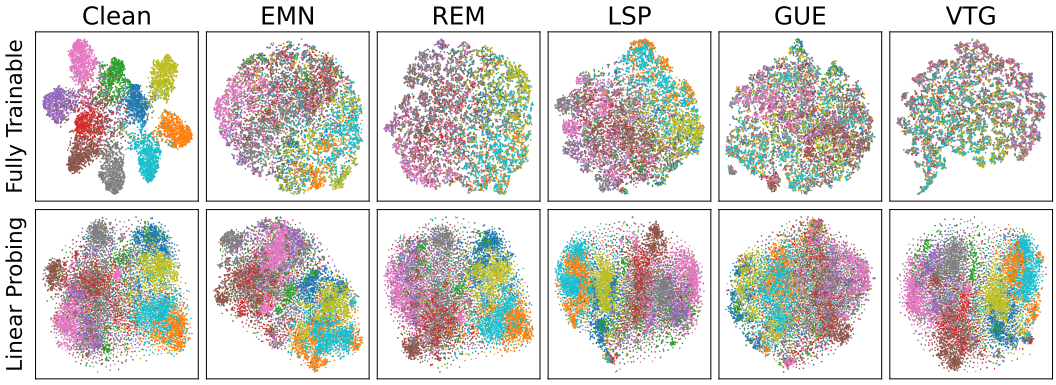


Figure 5: t-SNE visualization of clean test features under fully trainable and linear probing settings.

We provide additional t-SNE visualizations of clean test features in Figure 5, where the models are trained on perturbed data under either the fully trainable or the linear probing setting. As shown in the first row, existing UE methods generally achieve unlearnability by disrupting the model's feature representations on clean data, where features of different classes become heavily entangled and exhibit no clear class-wise separability. In contrast, under the linear probing setting (second row), the pretrained backbone remains frozen throughout training and only the linear head is updated. Consequently, the backbone is immune to perturbation-induced disruption and continues to produce discriminative features on clean test data. This fundamental discrepancy underscores the inherent challenge of injecting unlearnability under linear probing.

A.2 Details for Quantifying Semantic Shift

To quantify how each UE method alters samples in the feature space, we report two complementary metrics in Figure 1c: *Deviation from Original Prototypes* and *Shift toward the Semantic-Decoupled Subspace*. Both metrics are computed on the CIFAR-10 training set using an ImageNet-pretrained ResNet-18 backbone. The detailed computation procedure is described below.

For each clean image $\boldsymbol{x}$ and its perturbed version $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}$, we extract the corresponding features $\boldsymbol{z} = g(\boldsymbol{x})$ and $\boldsymbol{z}' = g(\boldsymbol{x}')$ with the frozen pretrained backbone. The two metrics also build upon several variables defined in Sec. 3.2, including the per-class prototype $\boldsymbol{c}_k$ (the mean clean features of class k), the global prototype $\boldsymbol{c}_g$ (the mean of all class prototypes), and the semantic basis $\mathbf{V}_{\text{sem}} \in \mathbb{R}^{D\times(K-1)}$ together with its orthogonal complement $\mathbf{V}_{\text{comp}} \in \mathbb{R}^{D\times(D-K+1)}$, both obtained by SVD on the centered semantic matrix $\mathbf{M} \in \mathbb{R}^{K\times D}$. Intuitively, $\mathbf{V}_{\text{sem}}$ spans the directions along with clean class clusters, while $\mathbf{V}_{\text{comp}}$ spans the semantic-decoupled subspace orthogonal to it.

Deviation from Original Prototypes. This metric measures how far the perturbed feature drifts from its corresponding clean class prototype. To eliminate dataset-level bias, we center both the perturbed feature $\boldsymbol{z}'$ and the target prototype $\boldsymbol{c}_k$ (class k) by subtracting the global prototype $\boldsymbol{c}_g$. The deviation is then computed as the cosine distance between the two centered vectors:

$$\mathcal{D}_{\text{deviation}} = 1 - \frac{(\boldsymbol{z}' - \boldsymbol{c}_g)^\top (\boldsymbol{c}_k - \boldsymbol{c}_g)}{\|\boldsymbol{z}' - \boldsymbol{c}_g\|_2 \, \|\boldsymbol{c}_k - \boldsymbol{c}_g\|_2}, \tag{11}$$

where larger values indicate greater dissimilarity between the perturbed sample and its ground-truth class prototype, reflecting a stronger deviation from the original semantic cluster. We reported the averaged values over all training samples in Figure 1c, where our method achieves the highest deviation among all baselines.

Shift toward the Semantic-Decoupled Subspace. This metric quantifies how strongly the perturbations push samples into the semantic-decoupled subspace. Let $\Delta\boldsymbol{z} = \boldsymbol{z}' - \boldsymbol{z}$ denote the feature

displacement induced by the perturbation. We measure this shift by projecting $\Delta \boldsymbol{z}$ onto $\mathbf{V}_{\text{comp}}$:

$$\mathcal{P}_{\text{shift}} = \left\| \mathbf{V}_{\text{comp}}^{\top} \Delta \mathbf{z} \right\|_2, \tag{12}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm. A larger $\mathcal{P}_{\text{shift}}$ indicates that the perturbation moves the sample further into the semantic-decoupled subspace. The reported values in Figure 1c are averaged over all training samples, where DIVERT attains the largest shift among all methods.

Overall, $\mathcal{D}_{\text{deviation}}$ captures how far a perturbed sample departs from its original semantics, while $\mathcal{P}_{\text{shift}}$ characterizes how much of this displacement falls into the semantic-decoupled subspace. Together, they show that DIVERT successfully pushes perturbed samples away from clean prototypes and toward the orthogonal subspace, empirically supporting the boundary-diversion mechanism.

A.3 Justification for the Rank of the Semantic Basis Matrix

In Section 3.2, we state that the centered semantic basis matrix $\mathbf{M} \in \mathbb{R}^{K \times D}$ has a rank of at most $K - 1$. This arises directly from the mean-centering operation during its construction.

Specifically, before applying ℓ_2 normalization, let the centered prototype for class k be denoted as $\tilde{\boldsymbol{c}}_k = \boldsymbol{c}_k - \boldsymbol{c}_g$. By definition, the global prototype $\boldsymbol{c}_g$ is the mean of all class prototypes. Consequently, the sum of all centered prototypes strictly equals the zero vector:

$$\sum_{k=1}^{K} \tilde{\boldsymbol{c}}_k = \sum_{k=1}^{K} \boldsymbol{c}_k - K \cdot \boldsymbol{c}_g = \mathbf{0}. \tag{13}$$

This linear dependence implies that one vector can always be expressed as a linear combination of the others. Therefore, the K centered prototypes can span a subspace of dimension at most $K - 1$.

In Eq. 3, the k-th row of $\mathbf{M}$ (denoted as $\mathbf{M}_k$) is obtained by scaling $\tilde{c}_k$ by a non-zero scalar factor $(1/\|\tilde{c}_k\|_2)$. Since scalar multiplication does not alter the geometric span of the vectors, the row space of $\mathbf{M}$ remains identical to the subspace spanned by $\{\tilde{c}_k\}_{k=1}^{K}$. Thus, the rank of $\mathbf{M}$ is inherently bounded by $K - 1$.

A.4 Choice of SVD for Semantic Decomposition

Table 9: Test accuracy (%) ↓ comparison between different decomposition strategies on CIFAR-10 with pretrained ResNet-18 under linear probing.

Decomposition	Test accuracy (%) ↓
QR	14.38
SVD (**Ours**)	**13.21**

We adopt SVD to obtain orthonormal bases for the class-prototype matrix and its orthogonal complement. The singular values characterize the strength of inter-class variation along the corresponding directions and help determine the numerical rank of the matrix, thereby identifying the informative components in high-dimensional features. This provides a convenient construction for separating the directions captured by the clean class prototypes from the complementary subspace used to form synthetic anchors. To examine the dependence on this decomposition, we replace SVD with a QR-based orthogonal-complement construction using the same class prototypes. As shown in Table 9, the QR-based variant achieves similar performance compared with the SVD-based construction. Both variants achieve low clean test accuracy, suggesting that effective protection is not specific to SVD and can also be achieved through an alternative construction of the semantic-decoupled subspace.

A.5 Empirical Evidence of Boundary Diversion

To examine whether the learned classifier relies on perturbation-induced separability that shifts the decision boundary, we train a linear head on the perturbed CIFAR-10 training set using a frozen ImageNet-pretrained ResNet-18. We then evaluate the same classifier on the perturbed training samples and their paired clean counterparts. As shown in Table 10, DIVERT achieves 99.99%

Table 10: Accuracy (%) on perturbed CIFAR-10 training samples and their paired clean counterparts with pretrained ResNet-18 under linear probing.

Method	Perturbed Samples ↑	Clean Counterparts ↓
EMN	96.05	44.83
LSP	80.50	67.94
Ours	**99.99**	**10.63**

accuracy on the perturbed samples but only 10.63% on their clean counterparts, yielding a larger accuracy gap than EMN and LSP. This discrepancy supports the classifier's reliance on perturbation-dependent separability and provides empirical evidence on the boundary-diversion mechanism.

A.6 Comparison with LSP

LSP [57] also builds on linearly separable perturbations. However, we note that DIVERT is inherently different from LSP in how it establishes linear separability and its relation to the clean semantic geometry. We discuss these differences below from three aspects.

Optimization Space. LSP constructs linearly separable perturbations in pixel space and relies on models trained from random initialization to learn the resulting shortcuts. However, pixel-space separability is not necessarily preserved by a pretrained backbone. DIVERT directly induces separability in the pretrained feature space by pushing perturbed features toward synthetic class anchors, making its optimization better aligned with the linear probing setting.

Semantic Dependence. LSP generates perturbations without access to the original data and does not explicitly account for the clean semantic geometry. Consequently, its construction does not ensure that the learned decision boundary deviates from this geometry. DIVERT uses clean class prototypes to decompose the feature space via SVD and employs cross-attention to construct synthetic anchors within the semantic-decoupled subspace. This design encourages the linear head to rely on spurious separability rather than the original prototype structure.

Empirical Comparison. The results in Table 1 show that DIVERT achieves stronger unlearnability than LSP under linear probing. The t-SNE visualizations in Figure 2 further provide a qualitative comparison of how the two methods affect feature distributions.

B More Experimental Details

B.1 Datasets

CIFAR-10/100. Both CIFAR-10 and CIFAR-100 [48] consist of 60,000 natural images at a resolution of 32×32 with 3 color channels, split into 50,000 samples for training and 10,000 for testing. They differ in label granularity, covering 10 and 100 object categories, respectively.

SVHN. The Street View House Numbers (SVHN) dataset [49] is collected from house number plates in Google Street View imagery. It comprises 73,257 images for training and 26,032 for evaluation, spanning 10 digit categories (0–9). Each sample is a 32×32 RGB crop centered on a single digit.

Tiny-ImageNet. Drawn from ImageNet, Tiny-ImageNet [50] covers 200 object categories. Each class provides 500 training samples and 50 validation samples, yielding 100,000 and 10,000 images, respectively. All images are resized to 64×64 with 3 color channels.

B.2 Implementation Details

During linear probing, all target models are pretrained on ImageNet and equipped with a randomly initialized linear head tailored to downstream datasets. For CNN-based architectures, we adopt the ImageNet-pretrained models provided by torchvision. For vision transformers, we employ Microsoft TinyViT [55]. Since the pretrained backbone is trained on the high-resolution ImageNet dataset and remains frozen, we upsample all downstream datasets to 224×224 resolution during linear probing. All baseline methods are reproduced at the same resolution to ensure a fair comparison.

C More Experiments

C.1 Evaluation on Fine-Grained Datasets

Table 11: Test accuracy (%) ↓ on the fine-grained Stanford Cars with pretrained ResNet-18 under linear probing.

Method	Test accuracy (%) ↓
Clean	46.62
EMN	28.26
LSP	35.32
Ours	**9.34**

To broaden our evaluation, we additionally assess DIVERT on Stanford Cars [72], a fine-grained classification dataset containing 16,185 images from 196 car categories. Recognizing these categories requires distinguishing subtle visual differences, providing a challenging complement to the coarse-grained benchmarks. We conduct the evaluation using ResNet-18 under linear probing. As shown in Table 11, DIVERT reduces clean-test accuracy from 46.62% to 9.34%. These results demonstrate that DIVERT provides the strongest protection among the evaluated baseline methods and remains effective in the fine-grained classification setting.

C.2 Protection without Ground-Truth Labels

Table 12: Test accuracy (%) ↓ on CIFAR-10 and CIFAR-100. DIVERT uses CLIP-based pseudo-labels throughout its protection pipeline.

Dataset	Clean	EMN	LSP	GUE	**Ours**
CIFAR-10	80.73	44.54	66.89	80.39	**15.69**
CIFAR-100	58.12	18.86	54.64	57.05	**4.95**

The main formulation of DIVERT uses class labels to construct prototypes, associate samples with anchors and class-wise perturbations, and optimize the surrogate classification objective. However, in many practical settings, label information is not available [73, 74, 75]. To extend our evaluation to settings without ground-truth labels, we extract features using a frozen CLIP ViT-B/32 image encoder, owing to its strong semantic alignment capability [76, 77, 78, 79], and apply K-means clustering to obtain pseudo-labels. These pseudo-labels replace ground-truth labels throughout the two-stage protection pipeline. As shown in Table 12, DIVERT achieves clean-test accuracies of 15.69% on CIFAR-10 and 4.95% on CIFAR-100, which yields the lowest accuracy among baseline methods. These results provide preliminary evidence that DIVERT remains effective with cluster-derived supervision, even though such pseudo-label supervision contains label noise [80, 81, 82], supporting its applicability when ground-truth labels are unavailable during perturbation generation.

C.3 Surrogate Model Sensitivity

Table 13: Sensitivity to surrogate architecture, pretraining dataset, and pretraining paradigm. We report clean test accuracy (%) ↓ under linear probing.

Surrogate	Pretraining dataset	Pretraining paradigm	**Ours**
ResNet-18	ImageNet-1K	Supervised	13.21
	Places365	Supervised	14.50
ResNet-50	ImageNet-1K	Supervised	14.29
	ImageNet-1K	Self-supervised (Barlow Twins)	15.90
	ImageNet-1K	Self-supervised (DINO)	15.93

We further evaluate the sensitivity of DIVERT to the surrogate model used for perturbation generation. The evaluated configurations cover different architectures, pretraining datasets (ImageNet-1K [1] and Places365 [62]), and pretraining paradigms, including supervised learning and self-supervised learning. As shown in Table 13, our method reduces the clean-test accuracy to nearly random-guessing levels across the evaluated configurations. The consistently low accuracies indicate that DIVERT retains effective protection beyond the default ImageNet-supervised surrogate.

C.4 Hyperparameter Sensitivity of Weight Coefficient

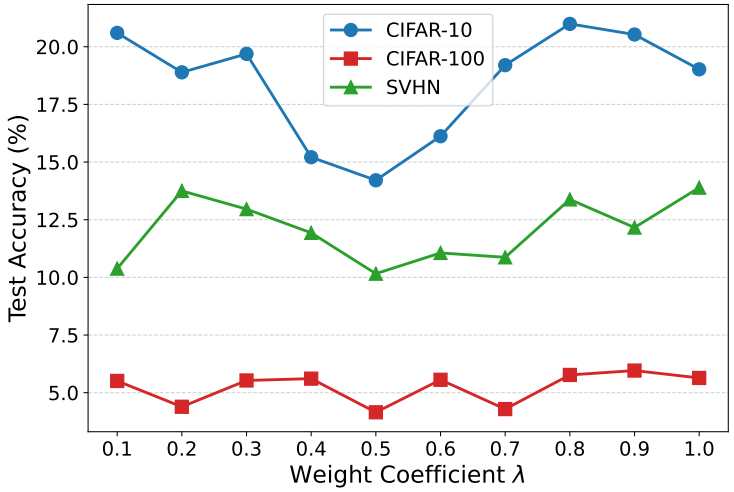


Figure 6: Effect of the weight coefficient λ on CIFAR-10, CIFAR-100, and SVHN.

We investigate the role of the weight coefficient λ in Eq. 10, which balances the diversity loss against the shifting loss and the cross-entropy loss. Figure 6 reports the clean test accuracy on CIFAR-10, CIFAR-100, and SVHN as λ varies from 0.1 to 1.0 on ResNet-18. Across the entire range, the accuracy fluctuates within a narrow band on all three datasets, demonstrating that our method is relatively insensitive to the selection of λ. We thus set λ to 0.5 as the default in all experiments.

C.5 Impact of Perturbation Budget

Table 14: Impact of the perturbation budget.

Dataset	Method	$2/255$	$4/255$	$8/255$	$16/255$
CIFAR-10	EMN	77.19	50.64	44.54	30.37
	Ours	**17.33**	**16.14**	**13.21**	**13.10**
CIFAR-100	EMN	56.45	42.15	18.86	15.51
	Ours	**32.73**	**10.30**	**3.21**	**2.94**

We examine the impact of perturbation magnitude by varying the ℓ_∞ budget from $2/255$ to $16/255$ on CIFAR-10 and CIFAR-100. As shown in Table 14, DIVERT consistently achieves lower clean-test accuracy than EMN at every tested budget on both datasets. We observe that increasing the perturbation budget generally strengthens protection. For example, on CIFAR-100, DIVERT reduces clean-test accuracy from 32.73% at $2/255$ to 3.21% at the default budget of $8/255$, while further increasing the budget to $16/255$ yields 2.94%. Similarly, on CIFAR-10, increasing the budget also enhances the unlearnability.

C.6 Complete JPEG Defense

We present the comprehensive evaluation of JPEG compression defenses under the linear probing protocol, as shown in Table 15. We observe a consistent decline in test accuracy across all models on

Table 15: Test accuracy (%) ↓ comparison against JPEG compression defense under linear probing with an ImageNet-pretrained ResNet-18 on CIFAR-10.

JPEG Compression	90	80	70	60	50	40	30	20	10
Clean	73.17	68.75	65.79	63.67	62.08	60.53	57.95	54.52	48.71
EMN	31.75	36.47	34.51	32.79	32.32	37.05	33.98	26.46	21.44
TUE	56.66	53.27	48.40	43.48	39.05	34.84	26.51	21.85	18.15
REM	59.47	52.08	57.58	52.77	47.21	41.17	32.24	26.18	22.07
LSP	58.78	50.03	40.74	37.04	32.17	26.60	23.71	20.88	20.72
VTG	69.18	65.12	60.46	54.51	49.72	41.68	32.19	25.13	21.32
Ours	**24.07**	**26.39**	**25.58**	**24.53**	**23.15**	**22.13**	**21.99**	**19.04**	**17.83**

both clean and different UE-protected datasets, as JPEG compression becomes more aggressive (quality decreasing from 90 to 10). This trend contrasts with prior findings in fully trainable settings [70], where stronger compression typically breaks unlearnable perturbations and subsequently restores test accuracy. We attribute this discrepancy to the inherent constraints of the linear probing paradigm. While aggressive compression continually undermines the crafted perturbations, it simultaneously corrupts the underlying semantic features of the images. Since the pretrained backbone remains frozen during linear probing, it cannot adapt to the new compressed images. For instance, the test accuracy on the clean baseline drops substantially from 73.17% to 48.71% as the JPEG quality decreases from 90 to 10. Consequently, the model is forced to rely on the degraded representations, which inherently bottlenecks the clean test accuracy and nullifies the potential gains from the removal of perturbations.

D More Analyses

D.1 Performance Gap between CNN and ViT

Table 16: Fisher Discriminant Ratio (FDR) across different pretrained backbones.

Pretrained Backbone	FDR
ResNet-18	0.158
ResNet-50	0.173
VGG-11	0.160
DenseNet-121	0.173
ViT	**0.239**

ViT generally exhibits better resistance to the perturbations, which results in relatively higher test accuracy, as shown in Table 1. To investigate the underlying mechanism, we analyze the class separability of clean features extracted by different pretrained backbones. Specifically, using the global and class-specific feature prototypes, we compute the inter-class scatter S_{inter} and intra-class scatter S_{intra}, which measure the dispersion of class prototypes around the global prototype and the dispersion of samples around their respective class prototypes, respectively. We then calculate the Fisher Discriminant Ratio (FDR) as $\text{FDR} = S_{\text{inter}}/S_{\text{intra}}$, where a higher FDR indicates greater inter-class separation relative to within-class dispersion.

As shown in Table 16, ViT exhibits a higher FDR compared with CNNs, which indicates stronger relative class separation in the evaluated ViT feature space. Under the same perturbation budget, such geometry may make it more difficult for perturbation-induced cues to dominate the clean semantic information and divert the decision boundary learned by the linear head. This observation offers an empirical explanation for the reduced protection effectiveness on ViT.

D.2 Mechanism of Cross-Architecture Transferability

The cross-architecture experiments in Table 4 demonstrate strong generalizability of the generated unlearnable perturbations. Inspired by the claim that although pretrained backbones differ in architec-

Table 17: Spearman correlation of inter-class prototype geometry compared with the pretrained ResNet-18 surrogate on CIFAR-10.

Target Backbone	Geometry Similarity
ResNet-18 (random initialization)	0.53
ResNet-50	0.98
VGG-11	0.97
DenseNet-121	0.95
GoogLeNet	0.98
ViT	0.91

ture, they tend to preserve similar relational geometry for the same downstream data [83, 84], we conduct experiments to investigate the geometric basis of cross-architecture transferability. Specifically, we compare their inter-class prototype geometry on CIFAR-10 using Spearman rank correlation, taking the pretrained ResNet-18 surrogate as the reference. This comparison measures the consistency in relative class-prototype relationships across feature spaces. For better comparison, we also include a randomly initialized ResNet-18 as a baseline.

As shown in Table 17, the pretrained target backbones exhibit high geometry similarity with the surrogate, with correlations ranging from 0.91 to 0.98. In comparison, the randomly initialized ResNet-18 achieves a correlation of only 0.53. These results indicate that pretrained backbones share a similar semantic geometry of classes despite their architectural differences. Since DIVERT constructs synthetic anchors with reference to class-prototype geometry, this shared semantic structure offers a plausible explanation for the observed transfer of protection beyond the surrogate backbone.

E Computational Cost

We evaluate the computational cost of DIVERT in terms of runtime, GPU usage, optimization convergence, and scaling. All experiments are conducted on a single NVIDIA L40S GPU.

Table 18: Perturbation optimization cost on CIFAR-10 using ResNet-18 with a batch size of 64.

Method	GPU Memory (GB)	Time (s / epoch)	Total Time (s)	Test Accuracy (%)
EMN	6.11	32.24	**515.93**	44.54
VTG	29.90	678.14	20344.17	70.82
Ours	**3.46**	**20.22**	1249.56	**13.21**

Runtime and GPU Usage. As shown in Table 18, DIVERT uses less GPU memory and requires less time per epoch than EMN. Compared with VTG, DIVERT reduces both memory usage and total runtime. It also achieves the lowest clean test accuracy among the compared methods, demonstrating stronger protection.

Table 19: Loss values during the two-stage optimization on CIFAR-10 with ResNet-18.

Epoch	1	15	30	45	60
$\mathcal{L}_{\text{shift}}$	0.370	0.189	0.142	0.114	0.095
$\mathcal{L}_{\text{div}}$	0.214	0.023	0.020	0.025	0.024
$\mathcal{L}_{\text{CE}}$	–	–	3.134	0.066	0.029

Optimization Convergence. Table 19 summarizes the loss trajectories during perturbation optimization. $\mathcal{L}_{\text{shift}}$ decreases steadily, while $\mathcal{L}_{\text{div}}$ drops early and subsequently remains low with minor fluctuations. The classification loss $\mathcal{L}_{\text{CE}}$ is introduced only in the second stage, during which it also decreases substantially. These trends indicate stable optimization across the two stages in the evaluated setting.

Table 20: Perturbation generation cost on Tiny-ImageNet subsets of varying sizes using ResNet-18.

Training Set Size	Time (s / epoch)	Total Time (s)
10,000 (10%)	5.45	382.52
25,000 (25%)	12.04	784.52
50,000 (50%)	23.02	1459.84
75,000 (75%)	33.50	2081.75
100,000 (100%)	44.68	2759.24

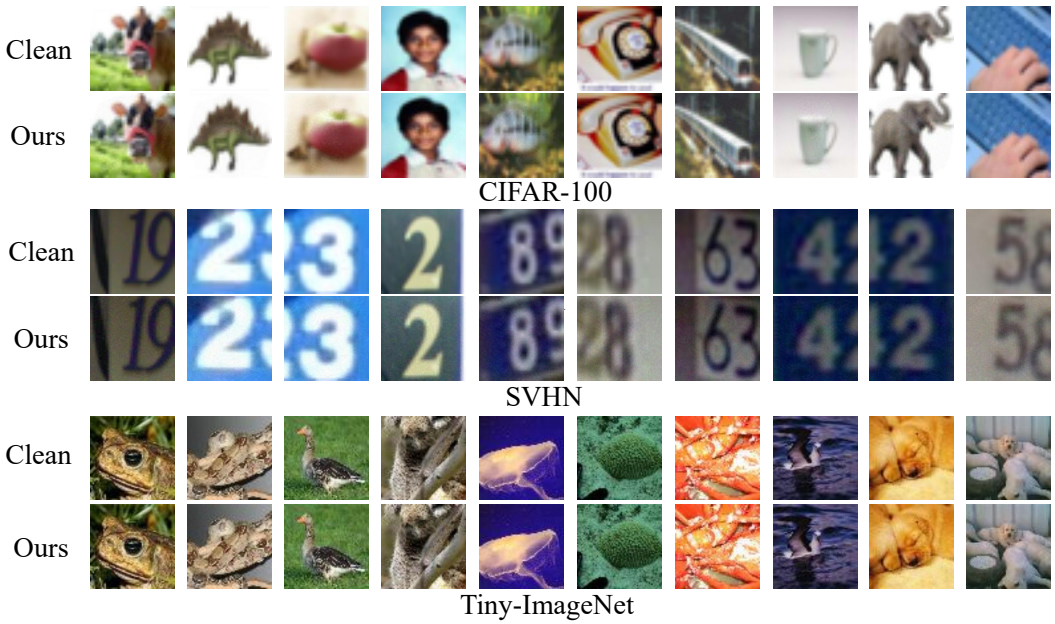


Figure 7: More visualizations of the perturbed images constructed by our method.

Scaling with Dataset Size. We evaluate scalability by varying the number of training samples on Tiny-ImageNet. As shown in Table 20, both per-epoch and total runtime grow approximately linearly with dataset size, indicating predictable computational scaling across the evaluated subsets.

F More Visualizations of Perturbed Images

To further demonstrate the imperceptibility of the crafted perturbations, we provide additional visualizations on CIFAR-100, SVHN, and Tiny-ImageNet, as illustrated in Figure 7. We observe that the perturbations remain visually unnoticeable across all datasets and do not impact the perceptual quality. These results further confirm that the generated perturbations maintain imperceptibility to human eyes while still effectively enforcing unlearnability against linear probing.

G Discussion on Potential Implications

DIVERT induces unlearnability by establishing spurious separability in a semantic-decoupled subspace and diverting the decision boundary. These design choices suggest two broader implications.

A New Design Perspective for UEs. Unlike UE methods that primarily aim to disrupt representation learning, DIVERT decomposes the pretrained feature space into a class-semantic subspace and its orthogonal complement, then deliberately establishes spurious class-wise separability. This highlights an alternative direction for UE design: even when a pretrained backbone is frozen and its clean features remain discriminative, protection can be achieved by diverting the decision boundaries learned by the downstream classifier.

Insight to Other Frozen-backbone Adaptation Paradigms. Beyond linear probing, many practical adaptation pipelines freeze a pretrained backbone and train only lightweight modules, such as classification heads, adapters, or multimodal connectors [85, 86, 87]. In these settings, protection strategies that rely on disrupting backbone learning may have limited effectiveness. The boundary-diversion principle may therefore guide the design of data-protection or watermarking methods that target downstream decision geometry, offering a potential direction for future research.

H Potential Societal Impact

This work adheres to the NeurIPS Code of Ethics. Our research investigates unlearnable examples as a data protection mechanism, with the goal of safeguarding individuals' data against unauthorized exploitation by machine learning systems. All experiments are conducted on publicly available datasets in compliance with their respective licenses, and our study involves neither human subjects nor sensitive personal data. Although techniques such as DIVERT could potentially be repurposed for adversarial ends, we underscore that the motivation behind this work is to advance privacy-preserving machine learning and strengthen defenses against unauthorized model training. We disclose our methodology and findings transparently to promote reproducibility and encourage responsible use within the research community.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We present the contributions and scope in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations in the conclusion section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the information required to reproduce the main results is included in the experiment settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code is available at `https://github.com/zhli-cs/DIVERT`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Relevant details are discussed in the experiment settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report results averaged over five runs with different random seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the compute resources in the implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The potential social impacts are discussed in the appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper aims at data protection, and there are no safety risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing assets are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code will be released after acceptance with detailed documents.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve these contents.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The experiments in this paper do not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLM is used only for polishing the writing. The core method development in this research does not involve LLMs as important components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.