WHEN PRIORS BACKFIRE: ON THE VULNERABILITY OF UNLEARNABLE EXAMPLES TO PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

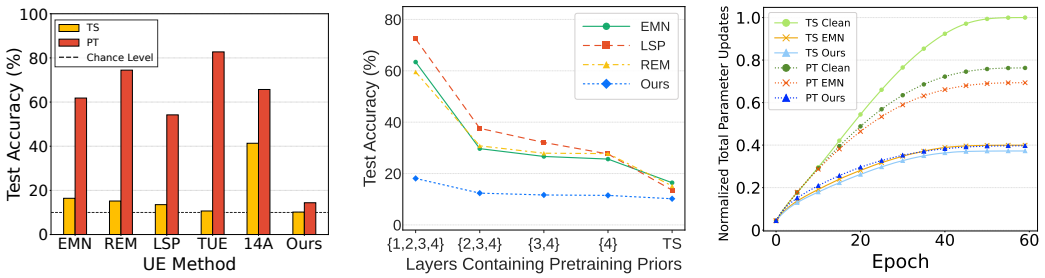
Unlearnable Examples (UEs) are introduced as a data protection strategy that generates imperceptible perturbations to mislead models into learning spurious correlations rather than real semantics. In this paper, we reveal a fundamental vulnerability of UEs that emerges when learning starts from a pretrained model. Specifically, our empirical analysis shows that even when data are protected by carefully crafted perturbations, pretraining priors still allow the model to bypass the shortcuts introduced by UEs and capture semantic information from the data, thereby nullifying unlearnability. To counter this effect, we propose **BAIT** (**B**inding **A**rtificial perturbations to **I**ncorrect **T**argets), a novel bi-level optimization formulation in which the inner level mirrors standard UE objectives, while the outer level enforces a dynamic association of perturbations with incorrect labels, deliberately misleading pretraining priors and preventing them from aligning with true semantics. This mislabel-perturbation binding mechanism blocks the pretrained model from readily establishing the true label-data relationship, so the learning process cannot quickly rely on image semantics and instead remains dependent on the perturbations. Extensive experiments on standard benchmarks and multiple pretrained backbones demonstrate that our approach produces UEs that remain effective in the presence of pretraining priors.

1 INTRODUCTION

The rapid growth of social media has resulted in a large amount of data shared online, enabling large-scale datasets (Deng et al., 2009; Lin et al., 2014) but also raising concerns over unauthorized usage. To safeguard privacy, Unlearnable Examples (UEs) have emerged as a promising data protection strategy. By injecting perturbations into training data, UEs mislead models into capturing synthetic shortcuts rather than underlying semantics, thereby degrading the performance on clean test data.

Existing UE studies (Huang et al., 2021; Wu et al., 2023) primarily target randomly initialized models. However, as training from scratch is annotation-intensive, practitioners commonly adopt pretrained backbones (Iofinova et al., 2022; Fang et al., 2023). Yet the behavior of UEs on pretrained models remains unexplored, leaving a critical gap in assessing their applicability in real-world deployment. Meanwhile, UEs essentially operate by injecting spurious shortcuts (Ren et al., 2022; Yu et al., 2022; Sandoval-Segura et al., 2022), and recent studies indicate that pretraining can improve robustness to spurious correlations (Tu et al., 2020; Izmailov et al., 2022; Mehta et al., 2022). This raises a natural question: **Can UEs remain effective when trained on pretrained models?**

In this paper, we uncover an overlooked yet fundamental vulnerability of UEs when applying to a pretrained model. As shown in Figure 1a, when trained on unlearnable perturbations, the test accuracies of pretrained models are substantially higher than their train-from-scratch counterparts, indicating that they circumvent the intended unlearnability and acquire meaningful semantics. We hypothesize that pretraining priors are the primary factor that enables models to bypass the spurious correlation introduced by UEs. To investigate the role of pretraining, we progressively remove priors by replacing the pretrained layers with randomly initialized ones and observe that the test accuracy consistently decreases, indicating that the protective effect of UEs strengthens as priors are reduced, as illustrated in Figure 1b. We further posit that the presence of pretraining priors facilitates a semantics–label pathway, which guides models to exploit underlying semantics rather than the perturbation-label shortcuts, thereby enabling the resumption of effective learning. To validate this,



(a) UE Performance Gap (b) Effect of Pretraining Priors (c) Parameter Updates in Training

Figure 1: Empirical analysis of the vulnerability of existing UE methods to pretraining priors. All experiments are conducted on CIFAR-10 with ResNet-18. (a) Existing UE methods, such as EMN (Huang et al., 2021), REM (Fu et al., 2022), LSP (Yu et al., 2022), TUE (Ren et al., 2022) and 14A (Chen et al., 2024), suffer severe unlearnability degradation when applied to pretrained (PT) backbones instead of train-from-scratch (TS) models. (b) We progressively replace layers of a four-layer ImageNet pretrained ResNet-18 with randomly initialized layers until obtaining a train-from-scratch model. The resistance to UEs steadily diminishes as pretraining priors are removed. (c) We report the normalized total parameter updates when training on clean data and UEs, with details in Appendix B.1. We observe that effective perturbations are capable of reducing parameter updates and misleading models into acquiring easy-to-learn shortcuts. In contrast, for existing UE methods such as EMN, pretrained models circumvent the spurious correlations and remain fully optimized.

we examine the parameter update dynamics when training on clean data and on UEs produced by EMN (Huang et al., 2021), as shown in Figure 1c. We observe that for pretrained models, the parameter update magnitude on UEs is comparable to that on clean data and considerably larger than training from scratch. This indicates that due to the semantic pathway within priors, the models are able to resume learning meaningful image features rather than being confined to the spurious correlations introduced by UEs. These findings highlight that existing UE methods are vulnerable to pretraining, as strong priors enable models to bypass their protection and acquire real semantics.

To address the aforementioned limitation, we propose **BAIT** (**B**inding **A**rtificial perturbations to **I**ncorrect **T**argets), a novel bi-level formulation designed to counteract priors and craft effective UEs against fully fine-tuned pretrained models. The core idea is to disrupt the data-label alignment reinforced by pretraining priors and instead recover the UE-induced spurious perturbation-label correlations. To achieve this, we bind perturbations to incorrect target labels that are semantically distinct from ground truth. Specifically, the inner level injects fixed perturbations that discourage the optimization of model parameters from encoding genuine semantics, while the outer level optimizes perturbations to map perturbed images onto designated incorrect labels, thereby misleading the underlying label-data relationships established by priors. This mislabel-perturbation binding mechanism blocks models from effortlessly exploiting pretraining priors, compelling reliance on perturbations and producing UEs that remain effective against pretrained backbones. As shown in Figure 1, our method successfully reduces the test accuracy to chance level against pretraining priors. Our main contributions are summarized as follows:

- We reveal that UEs suffer from a fundamental vulnerability when applied to pretrained backbones. We also empirically demonstrate that pretraining priors enable models to bypass UE-induced shortcuts and acquire true semantics.
- We propose BAIT, a bi-level optimization framework that binds perturbations to incorrect target labels, thereby disrupting the semantics-label mapping underpinned by pretraining priors and reconstructing the synthetic perturbation-label correlations.
- Extensive qualitative and quantitative experiments demonstrate the effectiveness and generalizability of BAIT against various pretrained backbones.

2 Related Work

Unlearnable Examples (UEs) have emerged as a promising protection approach to prevent unauthorized data usage. The goal is to inject imperceptible perturbations into data such that models trained

on these perturbed samples achieve high training accuracy but fail to generalize to the clean test set, with test accuracy dropping to the chance level. To achieve this, prior studies mainly introduce "shortcuts" during training, guiding models to memorize synthetic perturbation-label associations instead of learning real semantics (Huang et al., 2021; Sandoval-Segura et al., 2022; Wu et al., 2023; Wang et al., 2025). Recently, UEs have been extended to broader and more practical scenarios. For example, some works (Ren et al., 2022; He et al., 2023; Wang et al., 2024) investigate UEs in the context of unsupervised learning. Moreover, works like UC (Zhang et al., 2023) consider label-agnostic settings, and 14A (Chen et al., 2024) further explores cross-dataset transferability. These studies have greatly promoted the development of this literature.

However, existing efforts predominantly focus on applying UEs to train-from-scratch classifiers, which neglects the practical yet crucial case of pretrained models that underpin most modern applications, leaving a critical gap in the current UE literature. Some UE studies (Qin et al., 2023; Sun et al., 2024) incorporate pretrained models either during perturbation optimization or as an auxiliary testbed to showcase the effectiveness of their methods. However, they did not study how UEs perform when training starts from a pretrained model. In this paper, we make the first attempt to uncover the vulnerability of UEs against pretraining priors. To address this limitation, we design a novel bi-level optimization framework that optimizes perturbations to align perturbed samples with designated labels different from their original class. As a result, our approach prevents pretrained models from learning the underlying semantics-label pathway introduced by priors and forces them to rely on perturbations to establish spurious correlations between perturbations and labels.

3 Proposed Method

In this section, we first describe our UE setting that targets classifiers with pretraining priors. We then introduce BAIT, a novel bi-level optimization framework designed to bind perturbations to incorrect target labels that are distinct from the true semantics. Finally, we elaborate on the implementation details and optimization strategies of the proposed framework.

3.1 Problem Setup

In contrast to existing UE studies that target train-from-scratch classifiers, we introduce a new UE setting that focuses on rendering data unlearnable for pretrained classifiers. Such models possess rich prior knowledge from their pretraining data and thus are harder to be misled by the injected unlearnable perturbations. Below, we formulate the problem in the context of image classification.

Objectives. Let $(\boldsymbol{x}, y)$ be a labeled example, where $\boldsymbol{x} \in \mathcal{X}$ is a data point, and $y \in \mathcal{Y} = \{1, \dots, K\}$ is its label. We denote the clean training and test sets as $\mathcal{D}_c$, $\mathcal{D}_t$, respectively. The goal of UEs is to transform the training set $\mathcal{D}_c$ into an unlearnable dataset $\mathcal{D}_u$ by injecting perturbations $\boldsymbol{\delta}$ to each example as $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}$. Generally, the perturbation δ is constrained by $\|\boldsymbol{\delta}\|_p \leq \epsilon$, ensuring imperceptibility to human vision, where $\|\cdot\|_p$ denotes the L_p norm and ϵ is chosen to be sufficiently small. A pretrained classifier $f_\theta : \mathcal{X} \to \Delta_\mathcal{Y}$, where θ is the model parameters and $\Delta_\mathcal{Y}$ denotes the probability simplex over the label space $\mathcal{Y}$, is *fully fine-tuned* on $\mathcal{D}_u$. The effectiveness of perturbations $\boldsymbol{\delta}$ is then evaluated by measuring the performance degradation of f_θ on the clean test set $\mathcal{D}_t$, which would approximate the chance level if $\boldsymbol{\delta}$ can successfully counter the pretraining priors of f_θ and establish artificial perturbation-label correlations for the training process.

3.2 Binding Artificial Perturbations to Incorrect Targets

Existing works primarily focus on designing UEs to protect data from train-from-scratch classifiers. However, as shown in Figure 1, their unlearnability diminishes substantially when applied to pretrained backbones, where priors bypass spurious correlations and allow the model to learn real data-label relationships. This fundamental vulnerability severely limits the practical deployment of UEs in protecting personal data from unauthorized exploitation.

To address the aforementioned limitation, we craft BAIT, a novel bi-level optimization framework that aims at binding artificial perturbations to incorrect targets. BAIT binds perturbations to designated incorrect labels that are semantically different from the ground truth, deliberately steering learning away from genuine semantics. Concretely, we design a mislabel-perturbation binding

mechanism that impedes the inherent data-label alignment within priors and instead drives the model to rely on perturbations to align the perturbed samples with a designated incorrect target label. In this way, when samples are perturbed with different perturbations, the perturbed samples can be mapped toward different incorrect target labels y, regardless of the true class. As training proceeds, models are compelled to rely on perturbations rather than pretraining priors to make label predictions, thereby bypassing the strength of prior knowledge in building the intrinsic semantics-label pathway. Consequently, BAIT neutralizes the influence of pretraining priors and produces unlearnable examples that remain effective against pretrained backbones. The formulation of the bi-level BAIT framework is provided below.

$$\begin{aligned} \min_{\boldsymbol{\delta}} \quad & \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_c}\mathcal{L}\big(f_\theta(\boldsymbol{x}^i+\boldsymbol{\delta}^j),y^j\big) \\ \text{s.t.} \quad & \theta\in\arg\min_{\theta}\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_c}\mathcal{L}\big(f_\theta(\boldsymbol{x}^i+\boldsymbol{\delta}^i),y^i\big)\,, \end{aligned} \tag{1}$$

where $\boldsymbol{x}$, $\boldsymbol{\delta}$, and y are data samples, class-wise perturbations, and labels, respectively; superscripts i and j indicate class indices with $i,j\in\{1,\dots,K\}$, $i\neq j$, and K being the total number of classes; function f_θ denotes a pretrained surrogate model parameterized by θ; $\mathcal{L}$ is the cross-entropy loss. Our formulation is a min-min bi-level optimization problem with different goals at each level, where the inner level optimizes model parameters θ to align the input x^i with the ground truth label y^i by adding corresponding perturbation δ^i, then the outer level enforces cross-label assignment by adding perturbation δ^j to bind the same input x^i onto a designated incorrect label y^j.

Our BAIT framework generates perturbations that actively counteract pretraining priors. (1) The inner optimization updates model parameters with fixed unlearnable perturbations, where these perturbations assist in establishing spurious correlations between data and labels and keep the surrogate model from learning real image features. More precisely, a sample $\boldsymbol{x}$ from class i is perturbed by the corresponding perturbation $\boldsymbol{\delta}^i$ and is mapped to its ground-truth label y^i. (2) The outer optimization on perturbations is the core innovation that dynamically strengthens the intentional mislabel-perturbation binding. Unlike conventional UEs, where perturbations preserve the original data-label relationship ($\boldsymbol{x}^i+\boldsymbol{\delta}^i\rightarrow y^i$), we enforce cross-label assignment and explicitly bind the perturbed samples onto designated incorrect labels ($\boldsymbol{x}^i+\boldsymbol{\delta}^j\rightarrow y^j$), deliberately misleading pretraining priors and preventing alignment with true semantics. This mislabel-perturbation binding mechanism blocks the pretrained model from readily establishing true data-label correspondences with the guidance of priors, so that training is compelled to treat $\boldsymbol{\delta}$ as the dominant signal. As a result, when pretrained models are trained on these unlearnable examples, they are tricked by spurious perturbation-label associations rather than the underlying data-label relationship, thereby failing to capture true image semantics.

3.3 Strategy for Crafting Effective Unlearnable Examples

Perturbation Generation. We follow generator-based UE methods (Liu et al., 2024a; Chen et al., 2024) and develop a perturbation generator $\mathcal{G}$ to generate unlearnable perturbations. $\mathcal{G}$ is designed with a standard encoder-decoder structure, which takes raw images $\boldsymbol{x}$ as input and produces corresponding perturbations as $\boldsymbol{\delta}=\mathcal{G}(\boldsymbol{x})$. To ensure visual imperceptibility, we constrain the perturbation magnitude by $\|\boldsymbol{\delta}\|_\infty\leq\epsilon$ with $\epsilon=8/255$, in accordance with previous works (Huang et al., 2021; Ren et al., 2022; Liu et al., 2024a).

Learning to Learn Unlearnable Perturbations. Directly minimizing the full bi-level objective in Equation 1 is intractable. To overcome this challenge, we adopt a meta-learning strategy (Finn et al., 2017; Huang et al., 2020; Liu et al., 2024b), approximating the outer objective by unrolling the inner optimization for N steps. Specifically, at the n-th iteration, let θ_n denote the surrogate model weights and $\boldsymbol{\delta}$ denote the perturbation (initialized to zero at the start of training). We create a copy of the current surrogate weights, $\theta'_{n,0}\leftarrow\theta_n$, which is then updated through N inner-level optimization steps as follows:

$$\theta'_{n,m+1}=\theta'_{n,m}-\alpha\nabla_{\theta'_{n,m}}\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_c}\mathcal{L}\big(f_\theta(\boldsymbol{x}^i+\boldsymbol{\delta}^i),y^i\big)\,, \tag{2}$$

where $m\in\{0,1,\dots,N-1\}$, and α indicates the learning rate that controls the step size. This N-step unrolling enables a "look ahead" perspective during training, allowing us to explicitly assess

how perturbations introduced at the current step gradually influence the outer mislabel-perturbation binding objective after N inner updates.

The unrolled surrogate model with weights $\theta'_{n,N}$ is then employed in the outer-level optimization to update perturbations, enforcing the perturbed samples onto incorrect target labels and misleading the intrinsic semantics-label connection within pretraining priors, which is shown below:

$$\boldsymbol{\delta}_{n+1}^{j} = \boldsymbol{\delta}_{n}^{j} - \beta \nabla_{\boldsymbol{\delta}_{n}^{j}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_c} \mathcal{L}(f_\theta(\boldsymbol{x}^i + \boldsymbol{\delta}_{n}^{j}), y^j), \tag{3}$$

where $i, j \in \{1, \ldots, K\}$ with $i \neq j$ denote class indices; β indicates the learning rate to optimize perturbations. In this procedure, each sample $\boldsymbol{x}$ from class i is perturbed with $\boldsymbol{\delta}^j$ from a target class j and forced towards the target incorrect label y^j, which optimizes perturbations that counter pretraining priors and prevent the model from relying on priors to associate samples with real labels.

The updated perturbations are incorporated into the corresponding samples $\boldsymbol{x}$ at the inner level, and the surrogate model θ_n is then updated accordingly to θ_{n+1}, which serves as the initialization for the next iteration:

$$\theta_{n+1} = \theta_n - \alpha \nabla_{\theta_n} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_c} \mathcal{L}(f_\theta(\boldsymbol{x}^i + \boldsymbol{\delta}^i), y^i). \tag{4}$$

The above meta-learning procedures are executed iteratively throughout the entire training process, ultimately yielding the optimized perturbations.

Curriculum-Guided Target Label Selection. To enhance the effectiveness of perturbations by directing perturbed samples toward diverse target labels, we dynamically select target labels rather than relying on fixed incorrect targets. Drawing inspiration from curriculum learning (Bengio et al., 2009; Wang et al., 2021), we design a three-stage easy-to-hard strategy according to the likelihood of misclassification, as shown below.

- **Stage 1: Hard Negative Classes**. We first select the non-ground-truth class with the highest logit score. Since these classes are most easily confused with the true class, they provide a natural starting point for optimization.
- **Stage 2: Random Classes**. We then randomly select non-ground-truth classes as the incorrect target labels, increasing task difficulty and improving the generalizability of perturbations across varying target labels.
- **Stage 3: Most Dissimilar Classes**. Finally, we select the class with the lowest logit score. This constitutes the most challenging case, as perturbations must push predictions toward semantically unrelated classes.

This progressive curriculum learning strategy dynamically guides perturbations to perform mislabel-perturbation binding from easy to hard targets, thereby enhancing their ability to mislead pretraining priors away from aligning samples with true labels. As a result, models are forced to depend on these perturbations for classification throughout training.

4 Experiments

In this section, we evaluate the effectiveness of the proposed optimization framework against pre-trained backbones with fully trainable parameters on standard benchmarks. We further examine its transferability across different pretraining priors and network architectures. Moreover, we evaluate performance with train-from-scratch classifiers to demonstrate the effectiveness of our method in conventional UE settings. Finally, we provide qualitative visualizations to demonstrate the imperceptibility of perturbations.

4.1 Experimental Setup

Datasets and Backbones. The datasets include CIFAR-10 (Krizhevsky et al., 2009) and CIFAR-100 (Krizhevsky et al., 2009), which contain 50,000 training images and 10,000 test images, and SVHN (Netzer et al., 2011), which contains 73,257 training images and 26,032 test images. The evaluated ImageNet-pretrained backbone includes ResNet-18 (He et al., 2016), ResNet-50 (He et al., 2016), VGG-11 (Simonyan & Zisserman, 2014), DenseNet-121 (Huang et al., 2017), and

Table 1: Test accuracy (%) ↓ comparison between our method and baselines against **ImageNet-pretrained** backbones on CIFAR-10, CIFAR-100, and SVHN datasets.

Dataset	Backbone	Clean	EMN	TUE	REM	LSP	GUE	14A	**Ours**
CIFAR-10	ResNet-18	84.10	61.82	82.72	74.46	54.20	23.17	65.70	**14.40**
	ResNet-50	86.48	57.54	85.68	85.06	65.00	71.42	72.81	**14.82**
	VGG-11	88.39	80.90	87.67	63.32	55.63	45.77	50.80	**22.14**
	DenseNet-121	87.17	63.31	86.30	61.05	62.66	72.31	74.89	**20.32**
CIFAR-100	ResNet-18	55.73	55.53	55.20	55.22	36.98	54.09	33.67	**20.77**
	ResNet-50	60.73	45.78	59.54	58.69	42.84	53.14	36.58	**28.02**
	VGG-11	63.62	38.10	62.52	35.25	39.26	57.02	30.58	**26.64**
	DenseNet-121	61.92	49.94	61.14	48.37	36.27	56.78	44.34	**25.81**
SVHN	ResNet-18	90.96	37.55	41.15	76.45	38.91	39.68	81.47	**14.37**
	ResNet-50	92.09	30.29	35.52	87.38	45.48	91.74	85.59	**18.86**
	VGG-11	93.34	56.24	77.54	51.50	49.89	32.14	82.89	**17.75**
	DenseNet-121	92.97	58.22	40.53	71.87	53.56	73.34	86.83	**11.61**

GoogLeNet (Szegedy et al., 2015). Unless otherwise specified, evaluations are conducted on ImageNet-pretrained models with a learning rate of 0.001, in contrast to prior UE studies that rely on train-from-scratch classifiers. The original fully connected layer is modified to match the number of classes in each downstream dataset. We report classification accuracy on the clean test set as the evaluation metric, where lower accuracy reflects stronger unlearnability.

Baselines. We use six representative UE methods as baselines, which are Error-Minimizing Noise (EMN) (Huang et al., 2021), Transferable Unexample Examples (TUE) (Ren et al., 2022), Robust Error-Minimizing Noise (REM) (Fu et al., 2022), Linear Separable Perturbation (LSP) (Yu et al., 2022), Game Unlearnable Example (GUE) (Liu et al., 2024a), and Universal Perturbation Generator (14A) (Chen et al., 2024), and evaluate their unlearnable effect against pretrained classifiers.

Implementation Details. Our model is implemented in PyTorch and trained on a single NVIDIA A5000 GPU. If not specified otherwise, we utilize an ImageNet-pretrained ResNet-18 as the surrogate model and set its learning rate α to 0.1. We adopt the Adam optimizer (Kingma & Ba, 2015) with a learning rate β of 0.001 to optimize the perturbation generator. Perturbations are constructed in a class-wise fashion, where we first generate perturbations for individual samples and then compute their class-wise averages. The unrolling step is set to $N = 2$. For the three-stage curriculum learning strategy, we train each stage for 30 epochs on CIFAR-10 and CIFAR-100, and for 20 epochs on SVHN. Perturbations are bounded by $8/255$ for imperceptibility. All results are reported as averages over five runs with different random seeds.

4.2 Unlearnability under Pretraining Priors

Unlearnability against ImageNet-Pretrained Backbones. As previously discussed in Section 1, existing UE approaches are predominantly designed for train-from-scratch classifiers and exhibit a marked decline in effectiveness when transferred to pretrained backbones. In this work, we introduce BAIT to explicitly counter prior knowledge and generate perturbations that remain effective for pretrained backbones. As shown in Table 1, we conduct comprehensive evaluations against ImageNet-pretrained classifiers and compare our approach with representative baselines. Perturbations are optimized and evaluated on the same backbone to ensure consistency. The results show that our method outperforms current UE methods across all backbones and datasets. Specifically, it drives performance close to random guessing on CIFAR-10 and SVHN, and achieves substantial accuracy reductions on CIFAR-100 compared to baselines. Notably, even though 14A (Chen et al., 2024) incorporates a pretrained CLIP backbone to aid per-

Table 2: Test accuracy (%) ↓ comparison against a pretrained ResNet-18 with a pretrained surrogate model for perturbation optimization.

Method	CIFAR-10	CIFAR-100	SVHN
EMN*	59.84	45.20	32.33
TUE*	82.65	55.28	47.71
REM*	56.08	34.30	55.26
Ours	**14.40**	**20.77**	**14.37**

Table 3: Transferability across pretrained backbones with different priors. An ImageNet-pretrained ResNet-18 is used as the surrogate model for perturbation optimization, and the resulting unlearnable examples are evaluated against prior knowledge from CIFAR-10, CIFAR-100, and SVHN.

Dataset	Pretraining Prior	EMN	TUE	REM	LSP	GUE	**Ours**
CIFAR-10	CIFAR-100	48.58	82.26	60.88	60.64	27.05	**20.58**
	SVHN	28.33	49.00	37.91	38.53	15.55	**9.75**
CIFAR-100	CIFAR-10	33.67	61.05	24.77	35.08	66.93	**22.89**
	SVHN	62.57	62.18	16.77	26.14	59.10	**11.21**
SVHN	CIFAR-10	19.13	12.80	70.43	26.73	29.35	**11.27**
	CIFAR-100	28.75	19.47	64.42	47.71	24.14	**17.49**

Table 4: Cross-architecture test accuracy (%) $\downarrow$ comparison of our method and baselines on CIFAR-10. ResNet-18 is used as the surrogate model during perturbation generation, and the resulting unlearnable examples are evaluated on ImageNet-pretrained classifiers with diverse architectures to assess transferability across network structures.

Method	Network Architecture			
	ResNet-50	VGG-11	DenseNet-121	GoogLeNet
EMN	70.18	69.74	68.79	63.98
TUE	84.86	86.64	85.21	82.41
REM	83.36	76.19	82.45	79.67
GUE	26.94	27.62	27.78	25.69
Ours	**15.94**	**17.28**	**18.51**	**18.56**

turbation generation, our method surpasses it on all benchmarks. These results highlight the superior ability of BAIT to bind perturbations with designated incorrect labels, compelling models to rely on perturbations rather than priors and thereby preventing models from learning genuine semantics.

Further Comparison to Re-implemented Baselines with ImageNet-Pretrained Priors. Several UE methods generally do not exploit pretraining priors when generating perturbations. Therefore, a natural question is whether their limitations arise from this omission, and whether their performance would improve if assisted with such priors. To rigorously examine this, we re-implement representative surrogate-based UE methods, replacing their randomly initialized surrogates with ImageNet-pretrained models to enable a fairer and more comprehensive comparison. The revised variants are denoted as EMN*, TUE*, and REM*, respectively. As shown in Table 2, even with access to ImageNet priors during perturbation optimization, our method consistently outperforms baselines, exhibiting stronger capabilities to erode and disrupt pretrained knowledge. This advantage enables our unlearnable examples to reliably counter pretrained backbones in practical applications and to provide sustained protection against unauthorized data usage.

Unlearnability across Different Pretraining Priors. As introduced in Section 3.2, we employ an ImageNet-pretrained surrogate model during the perturbation optimization and evaluate the resulting unlearnable examples on corresponding ImageNet-pretrained backbones. To consider practical scenarios where backbones may be pretrained on different datasets, we further examine the generalizability of BAIT to other pretraining priors derived from CIFAR-10, CIFAR-100, and SVHN. As shown in Table 3, our method consistently outperforms baselines in this transferability-focused scenario, demonstrating its ability to bypass priors, thereby generating unlearnable examples effective across diverse pretrained backbones. Moreover, we observe that backbones pretrained on datasets semantically similar to the perturbed one exhibit stronger resistance to UEs. For example, when the perturbed dataset is CIFAR-10, baselines show greater resistance with CIFAR-100 priors (daily objects) than with SVHN priors (digits). This indicates that broader and more semantically aligned priors provide greater resistance to UEs, thereby further highlighting the superiority of our approach in Table 1, where extensive ImageNet-pretrained priors are considered.

Unlearnability across Different Architectures We further examine the transferability of our perturbations across different network architectures. As shown in Table 4, perturbations are optimized with ResNet-18 as the surrogate model and evaluated on various backbones, including ResNet-

Table 5: Impact of each stage of curriculum-guided target label selection against an ImageNet-pretrained ResNet-18 model. "Stage1", "Stage2", "Stage3" denotes selecting target labels from "Hard Negative Classes", "Random Classes", and "Most Dissimilar Classes", respectively.

Stage1	Stage2	Stage3	CIFAR-10	CIFAR-100	SVHN
✓			23.68	35.75	21.75
✓	✓		20.60	23.94	16.58
✓	✓	✓	**14.40**	**20.77**	**14.37**

Table 6: Test accuracy (%) ↓ comparison between our method and baselines against standard **train-from-scratch** classifiers on CIFAR-10.

Method	ResNet-18	ResNet-50	VGG-11	DenseNet-121
EMN	16.42	13.45	16.93	14.71
TUE	10.67	12.23	12.79	17.60
REM	15.18	25.57	47.68	41.54
LSP	13.54	15.94	17.15	17.47
GUE	13.25	12.97	23.89	13.71
14A	41.34	24.85	39.87	48.97
Ours	**10.18**	**10.01**	**10.13**	**10.05**

50 (He et al., 2016), VGG-11 (Simonyan & Zisserman, 2014), DenseNet-121 (Huang et al., 2017), and GoogLeNet (Szegedy et al., 2015). The results demonstrate that our method induces unlearnability across all target architectures, driving test accuracy close to chance level. Compared with baselines, it consistently achieves superior performance, underscoring its strong cross-architecture transferability and practical effectiveness.

4.3 Ablations and Further Analyses

Impact of Curriculum-guided Target Label Selection. The selection of incorrect target labels plays a crucial role in optimizing perturbations to establish effective mislabel-perturbation binding, and we conduct ablation studies to assess its role. As shown in Table 5, unlearnability against pre-trained priors improves progressively with the inclusion of each stage. Specifically, we observe that selecting hard negative classes as the incorrect target labels (Stage 1) introduces the basic unlearnable effect, while incorporating random classes (Stage 2) and the most dissimilar classes (Stage 3) further strengthen it. These results highlight the contribution of each curriculum-guided stage and demonstrate that the synergistic integration of all three stages leads to the strongest performance.

Unlearnability against Conventional Train-from-Scratch Backbones. We argue that the mislabel-perturbation binding mechanism of the proposed BAIT also provides a feasible solution for train-from-scratch classifiers. To assess its effectiveness, we conduct conventional UE evaluations as shown in Table 6. We observe that while most baseline methods achieve satisfactory results and introduce unlearnability for randomly initialized models, our method outperforms them and achieves lower test accuracy. These results highlight the broad applicability of BAIT to both pre-trained backbones and train-from-scratch backbones.

Visualization of Perturbed Samples. The imperceptibility of perturbations to human vision serves as a key evaluation criterion for UEs. Therefore, we visualize our perturbed images on CIFAR-10 and make comparisons with ex-

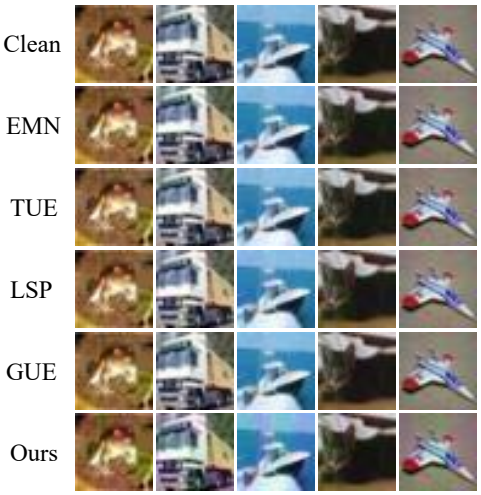


Figure 2: Perturbed examples on CIFAR-10.

Table 7: Test accuracy (%) ↓ comparison against ImageNet-pretrained backbones on larger datasets, including Flowers and ImageNet subset. Note that ImageNet* and ImageNet† denote two randomly selected 100-class subsets of ImageNet with no overlapping classes.

Downstream Dataset	Pretraining Prior	EMN	TUE	REM	LSP	GUE	**Ours**
Flowers	ImageNet	44.36	46.63	43.91	47.76	45.18	**24.91**
ImageNet*	ImageNet†	51.98	59.00	59.12	59.44	37.72	**24.22**

Table 8: Test accuracy (%) ↓ comparison against an ImageNet-pretrained ResNet-18 between our method and baselines under defenses, where CIFAR-10 is utilized as the downstream dataset. "Ortho. Proj." denotes the Orthogonal Projection Defense (Sandoval-Segura et al., 2023).

Method	w/o	Cutout	CutMix	Mixup	Ortho. Proj.
Clean	84.10	83.25	83.06	84.14	74.55
EMN	61.82	51.60	54.71	68.70	43.89
TUE	82.72	81.54	81.35	82.69	70.47
REM	74.46	65.39	73.71	78.08	67.38
LSP	54.20	44.68	51.53	59.44	55.06
GUE	23.17	29.31	28.30	34.48	49.37
Ours	**14.40**	**13.69**	**14.58**	**21.08**	**29.11**

isting UE methods. As illustrated in Figure 2, we observe that the perturbations optimized by BAIT are generally invisible to humans. We also provide more visualizations in Appendix E.1.

Evaluation on Larger Datasets. We add additional results on Flowers (Nilsback & Zisserman, 2008) and ImageNet to further demonstrate the effectiveness of our method on larger datasets. For Flowers, we evaluate our perturbations against the standard ImageNet-pretrained ResNet-18. For ImageNet, to avoid overlap between the pretraining and fine-tuning data since both originate from ImageNet, we randomly select 100 classes from the full 1000-class ImageNet as the downstream fine-tuning data, denoted as ImageNet*. We then randomly select another 100 classes, with no overlap with ImageNet*, to pretrain a ResNet-18 model and use it during evaluation. This pretraining dataset is denoted as ImageNet†. As shown below in Table 7, our method outperforms baselines with a clear margin, highlighting its effectiveness on larger datasets.

Resistance to Standard Defense Strategies. To evaluate the efficacy against both pretraining priors and defense strategies, we train models on perturbed data with standard data transformations, including Cutout (DeVries & Taylor, 2017), CutMix (Yun et al., 2019), and Mixup (Zhang et al., 2018), as shown in Table 8. We find that these transformations exert little influence on the unlearnability against pretrained backbones. We attribute this to the fact that pretraining priors already function as a strong defense, enabling models to bypass the spurious correlations introduced by perturbations. Nevertheless, our method remains effective and consistently outperforms baselines, underscoring its superiority in countering pretraining priors even in the presence of additional defenses.

Resistance to Advanced Defense Strategies. We further evaluate our method with advanced defenses, including JPEG compression (Liu et al., 2023) and Orthogonal Projection defense (Sandoval-Segura et al., 2023). JPEG compression removes high-frequency components and introduces frequency artifacts, making it a challenging perturbation scenario for unlearnable examples. Orthogonal Projection defense aims at learning a linear model, then projects UEs to be orthogonal to linear weights, which removes the most predictive dimensions of the data. Both evaluations are conducted on the CIFAR-10 dataset with an ImageNet-pretrained ResNet-18 model. As shown in Table 8, we observe that our method exhibits remarkable performance compared to comparative baselines when applying orthogonal projection defenses. Moreover, as shown in Table 9, our method outperforms baseline methods across all JPEG defenses with different compression quality levels 90, 80, and 70, etc. We observe that our method reduces clean test accuracy greatly and approximates random guessing level until the JPEG quality is set to 10, while baseline methods fail to introduce unlearnability even when the JPEG quality is set to higher values like 90. Considering images would be distorted visibly with a strong JPEG compression quality of 10, we argue that our method exhibits exceptional resistance against the JPEG defense.

Table 9: Test accuracy (%) ↓ comparison against an ImageNet-pretrained ResNet-18 with JPEG compression defense on CIFAR-10.

JPEG Compression	90	80	70	60	50	40	30	20	10
Clean	82.64	82.13	81.08	80.58	80.03	79.19	77.77	75.64	73.95
EMN	68.22	69.09	69.23	69.13	69.91	71.32	73.47	74.18	74.46
TUE	81.21	80.41	79.91	79.62	79.70	79.26	78.91	77.81	75.11
REM	77.15	77.29	78.08	78.17	78.34	78.55	78.33	78.14	75.33
LSP	55.47	57.24	59.56	61.43	63.48	63.84	64.86	68.53	71.00
GUE	60.94	65.59	66.73	66.86	68.39	69.18	70.11	72.33	72.88
Ours	**15.91**	**16.47**	**16.70**	**16.96**	**17.12**	**18.93**	**20.41**	**27.48**	**68.08**

4.4 Qualitative Analysis on the Optimization Process of BAIT

To provide further qualitative analyses regarding the optimization of perturbations, we use two variants of CIFAR-10 training samples to test the surrogate model, namely the original clean training samples and the perturbed training samples. During the optimization of BAIT, we plot the accuracy curve of the pretrained surrogate model f_θ on both types of samples, as illustrated in Figure 3. At the early stage of optimization, the surrogate model benefits from its inherent pretraining priors and acquires meaningful semantics, as reflected by the simultaneous improvement of performance on both perturbed and clean training images. However, as the optimization of perturbations, a divergence emerges: the accuracy on perturbed images keeps rising, while the accuracy on the corresponding clean images progressively declines toward chance level. This indicates that perturbations gradually mislead the real data-label association guided by pretraining priors and force the surrogate to rely on the spurious correlations between perturbations and labels, which prevents models from learning real semantics and further demonstrates the effectiveness of BAIT.

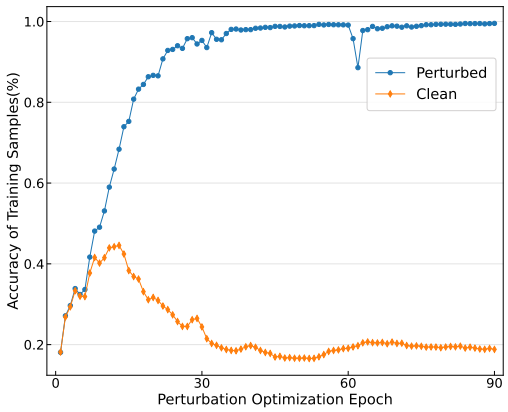


Figure 3: Training accuracy curve on CIFAR-10, illustrating that BAIT successfully misleads the ImageNet-pretrained surrogate model during perturbation optimization.

5 Conclusion

In this paper, we reveal an overlooked yet fundamental vulnerability of unlearnable examples that emerges when training starts from a pretrained backbone. We empirically show that the semantics-label pathway derived from pretraining priors enables pretrained models to bypass the shortcuts injected by UEs and acquire genuine semantics, even when data are protected by carefully crafted perturbations. To counter these priors and construct UEs that remain effective against pretrained backbones, we propose BAIT (Binding Artificial perturbations to Incorrect Targets), a novel bi-level optimization framework. Specifically, the inner level injects perturbations to keep models from learning the underlying real semantics, while the outer level enforces perturbed samples to align with designated incorrect target labels, thereby establishing mislabel-perturbation bindings that mislead pretraining priors and compel the model to rely on perturbations and the reconstructed spurious perturbation-label correlations during training. To further stabilize optimization, we incorporate meta-learning and dynamically select the target incorrect labels with a curriculum learning strategy. Extensive experiments across diverse backbones demonstrate that BAIT achieves superior performance over state-of-the-art methods.

Limitations. This paper investigates the vulnerability of UEs against pretrained models within image classification and designs a novel bi-level optimization framework. However, the transferability from classification to other downstream tasks (e.g., segmentation) remains limited, as discussed in Appendix F.3. Given the broad potential of cross-task transferability, we recognize this as an important research direction and plan to explore it in future work.

Ethical Statement

This work is conducted in accordance with the ICLR Code of Ethics. Our study focuses on the design of unlearnable examples as a data protection mechanism, aiming to safeguard individuals' data from unauthorized exploitation in machine learning systems. All experiments are carried out on publicly available datasets under their respective licenses, with no involvement of human subjects or sensitive personal data. While techniques such as BAIT could, in principle, be misused for adversarial purposes, we emphasize that our intent is to strengthen privacy-preserving learning and to advance defenses against unauthorized model training. We report our methods and findings transparently to foster reproducibility and responsible use.

Reproducibility Statement

We have taken several measures to ensure the reproducibility of our work. All datasets used in our experiments are publicly available and described in detail in the main text. The implementation of the proposed BAIT framework, including model architectures, training schedules, and evaluation procedures, is fully documented, and the source code is provided in the supplementary materials. Furthermore, additional hyperparameter settings are included in the implementation details to further support reproducibility.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, pp. 41–48, 2009.

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*, pp. 7700 – 7711, 2024.

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*, pp. 248–255. Ieee, 2009.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

Alex Fang, Simon Kornblith, and Ludwig Schmidt. Does progress on imagenet transfer to real-world datasets? *Advances in Neural Information Processing Systems*, 36:25050–25080, 2023.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

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.

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

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*, pp. 770–778, 2016.

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*, pp. 4700–4708, 2017.

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.

W Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, and Tom Goldstein. Metapoison: Practical general-purpose clean-label data poisoning. *Advances in Neural Information Processing Systems*, 33:12080–12091, 2020.

Eugenia Iofinova, Alexandra Peste, Mark Kurtz, and Dan Alistarh. How well do sparse imagenet models transfer? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12266–12276, 2022.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35: 38516–38532, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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

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

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

Yixin Liu, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, and Lichao Sun. Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 24219–24228, 2024b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

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

Raghav Mehta, Vítor Albiero, Li Chen, Ivan Evtimov, Tamar Glaser, Zhiheng Li, and Tal Hassner. You only need a good embeddings extractor to fix spurious correlations. *arXiv preprint arXiv:2212.06254*, 2022.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin 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, pp. 4. Granada, 2011.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.

Tianrui Qin, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. Learning the unlearnable: Adversarial augmentations suppress unlearnable example attacks. *arXiv preprint arXiv:2303.15127*, 2023.

Jie Ren, Han Xu, Yuxuan Wan, Xingjun Ma, Lichao Sun, and Jiliang Tang. Transferable unlearnable examples. *arXiv preprint arXiv:2210.10114*, 2022.

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*, pp. 3862–3871, 2023.

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.

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.

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

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.

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*, pp. 1–9, 2015.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633, 2020.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

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.

Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2021.

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.

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*, pp. 68–85. Springer, 2022.

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

Kai Ye, Liangcai Su, and Chenxiong Qian. How Far Are We from True Unlearnability? In *The Thirteenth International Conference on Learning Representations*, 2025.

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*, pp. 2367–2376, 2022.

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*, pp. 6023–6032, 2019.

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

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*, pp. 3984–3993, 2023.

APPENDIX

This appendix provides more comprehensive analyses of the vulnerability of UEs against fully-tuned pretrained backbones and our proposed bi-level min-min optimization framework. It expands upon the main paper by incorporating additional architectural details, extended empirical analyses, qualitative visualizations, and additional verification experiments, thereby offering a more in-depth and more holistic understanding of the mechanism underlying our approach and the broader implications for unlearnability research.

A THE USE OF LARGE LANGUAGE MODELS (LLMS).

We declare that Large Language Models were used exclusively for language polishing, including correcting grammar, improving readability, and ensuring consistency in terminology. All core ideas, empirical demonstration, experimental design, and result analyses were entirely conceived and conducted by the authors.

B DISCUSSION ON PARAMETER UPDATES AND SEMANTIC LEARNING

B.1 DETAILS OF THE PARAMETER UPDATES ILLUSTRATION

In Figure 1c, we present the total parameter updates when models are training on clean data and unlearnable data crafted by EMN and our method. We use the global maximum normalization to process the parameter update data, with details shown below.

Given the six groups of parameter changes measured during training, we use the L2 norm to calculate the parameter updates between epochs and denote the cumulative change of group k at epoch t as

$$v_t^{(k)} = \sum_{i=1}^{t} \|\Delta\theta_i^{(k)}\|_2, \tag{5}$$

where $\Delta\theta_i^{(k)}$ represents the parameter update at epoch i for group k. Since $v_t^{(k)}$ is monotonically non-decreasing with respect to t, its maximum is attained at the final epoch T. Therefore, the global normalization factor is defined as

$$M = \max_{1\le k\le 6} v_T^{(k)}. \tag{6}$$

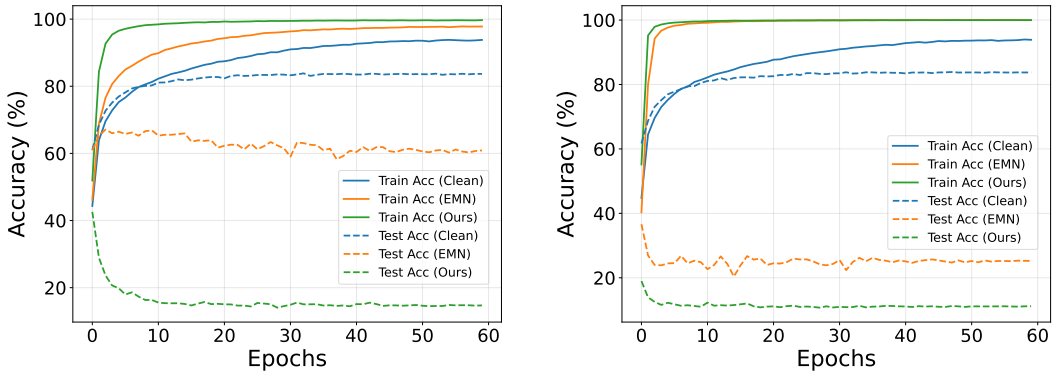
The normalized cumulative change is then computed as

$$\tilde{v}_t^{(k)} = \frac{v_t^{(k)}}{M}, \quad \forall\, t = 1, \ldots, T,\ k = 1, \ldots, 6. \tag{7}$$

This normalization ensures that all curves are scaled into the interval $[0, 1]$, facilitating a fair comparison of their relative growth dynamics during training.

B.2 RELATIONSHIP BETWEEN PARAMETER UPDATES AND SEMANTIC LEARNING

To further examine the relationship among parameter update magnitude, training duration, and model performance (Ye et al., 2025), and to validate that effective UEs can reduce parameter updates and limit the acquisition of semantics, we illustrate the training accuracy and testing accuracy curves when pretrained models are fine-tuned on clean data and unlearnable data, as shown in Figure 4. We are particularly interested in examining how the accuracy curve evolves when fine-tuning a pretrained model on UEs. As illustrated in Figure 4a, we observe that as the training accuracy of EMN improves, the testing accuracy maintains a high level of more than 60%, showing that the model fine-tuned on EMN-crafted unlearnable data actually learns real semantics. We also observe a similar phenomenon with a randomly initialized model in Figure 4b. This is consistent with our empirical findings in Figure 1c, where the parameter updates of EMN are larger than those of our method and close to clean training. As for the accuracy curve of our method, the testing accuracy is actually decreasing to chance level as the training accuracy increases, demonstrating that models are not learning meaningful semantic knowledge. The parameter updates curve of our method in



(a) Training starts from a pretrained model.

(b) Training starts from a randomly initialized model.

Figure 4: Accuracy curves illustration during UE evaluation on CIFAR-10 with ResNet-18.

Table 10: The architecture of the perturbation generator.

Block Name	Layer	Number
Down-sampling layers		
Conv	Conv (3×3) InstanceNorm ReLU	$\times 6$
Bottleneck layers		
Residual	ReflectionPad Conv (3×3) BatchNorm ReLU ReflectionPad Conv (3×3) BatchNorm	$\times 8$
Up-sampling layers		
ConvTranspose	ConvTranspose (3×3) InstanceNorm ReLU	$\times 5$
ConvTranspose	ConvTranspose (6×6) Tanh	$\times 1$

Figure 1c also stands for this claim, where it has fewer parameter updates compared to EMN and clean training.

C The Architecture of Perturbation Generator

In the main paper, we devise a versatile transferable generator to produce transferable perturbations and craft unlearnable examples. We denote our perturbation generator as $\mathcal{G}$, which employs a standard encoder-decoder architecture. This structure comprises three down-sampling convolution layers, four residual blocks (He et al., 2016), and three transposed convolution layers. The detailed architecture is shown in Table 10.

D Further Analysis on Pretraining Priors against Unlearnability

D.1 Impact of Pretraining Priors against Unlearnability

To further examine the role of pretraining priors in unlearnability, we freeze individual layers of an ImageNet-pretrained ResNet and analyze their impact on test accuracy. As shown in Figure 5, freez-

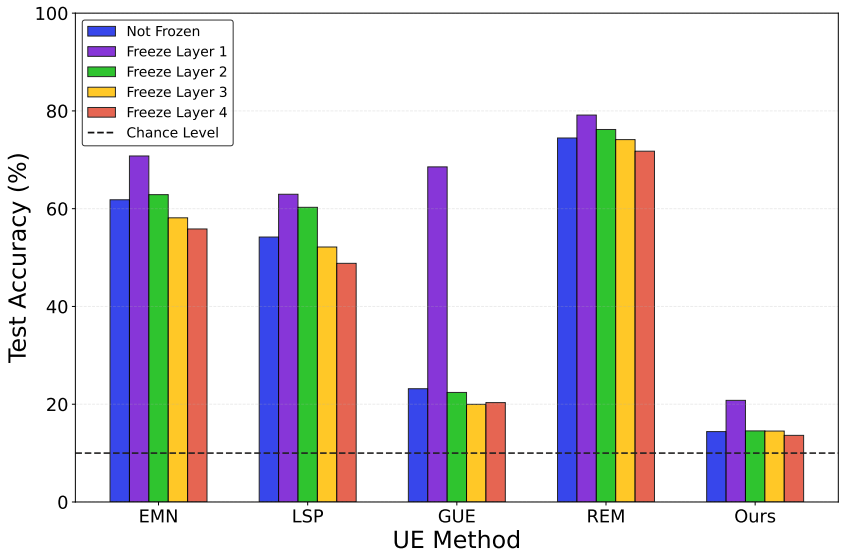


Figure 5: Effect of each pretraining layer of ResNet-18 against unlearnability on CIFAR-10.

Table 11: Test accuracy change (ΔAcc) on the original 1000-class ImageNet validation set when the ImageNet-pretrained model is fine-tuned on CIFAR-10.

Fine-tuning Dataset	Accuracy Variation (ΔAcc)			
	Epoch 1	Epoch 2	$\cdots$	Epoch 60
Clean CIFAR-10	-36.45	-53.83	$\cdots$	-55.28
Unlearnable CIFAR-10 (Ours)	-40.66	-55.11	$\cdots$	-55.35

ing layer 1 or layer 2 leads to an increase in test accuracy for several UE methods. This suggests that when low-level priors, which typically encode pixels, edges, and textures, are preserved from disruption by UEs, models are able to acquire more genuine semantics from the data. Notably, GUE exhibits a pronounced accuracy surge when the first layer is frozen, indicating that its perturbations mainly target shallow priors and thus become ineffective once these priors are fixed. In contrast, our method remains relatively effective even when shallow layers are frozen. Furthermore, freezing deeper layers, such as layer 3 and layer 4, results in a decrease in accuracy, suggesting that perturbations are now having no choice but to disrupt shallow priors and consequently achieve stronger unlearnability compared to the fully trainable pretrained scenario.

D.2 Performance Variation on Original Pretraining Data during Fine-tuning

To discuss the relationship between pretraining data and fine-tuning data, we conduct additional experiments to examine the performance variation on the original data during fine-tuning on the downstream new data. We note that, due to catastrophic forgetting, fine-tuning a pretrained model on downstream data is likely to cause a sharp decline in performance on the original data. To display this reduction more clearly, we report the relative performance variation when models are fine-tuned on either clean downstream data or our UEs. Specifically, we fine-tune the pretrained model on CIFAR-10 while monitoring the variation in its performance on the original 1000-class ImageNet dataset compared to its performance before fine-tuning. As shown in Table 11, we observe that the accuracy reduction is similar for both clean and UE fine-tuning data. This indicates that the observed performance decrease arises from catastrophic forgetting rather than from the unlearnable perturbations. Therefore, we argue that our perturbations effectively protect downstream user data without compromising the model's overall capability or increasing the risk of catastrophic forgetting during fine-tuning.

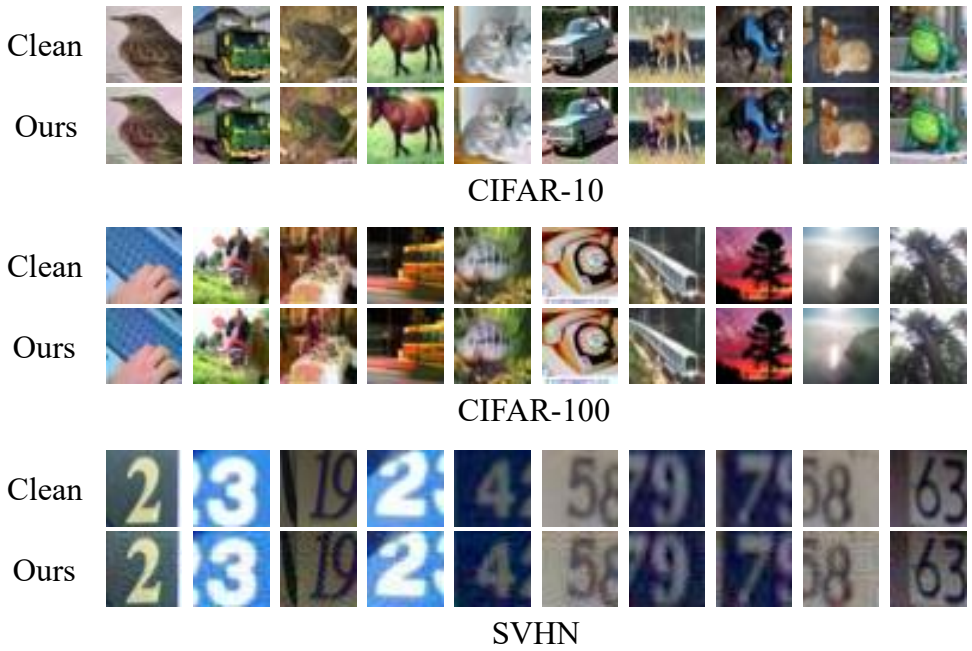


Figure 6: More visualizations of the unlearnable examples crafted by BAIT on CIFAR-10, CIFAR-100, and SVHN.

E More Visualizations

E.1 Visualizations on Unlearnable Examples

To further ensure the imperceptibility of our unlearnable examples, we provide additional visualizations on CIFAR-10, CIFAR-100, and SVHN, as illustrated in Figure 6. Across all datasets, the perturbations remain visually subtle, generally exhibiting small magnitudes that do not degrade the perceptual quality of the images. These results further confirm that the crafted perturbations maintain imperceptibility to human observers while still enforcing unlearnability on models.

E.2 T-SNE Visualizations

To illustrate the superior effectiveness of our BAIT, we further present the t-SNE visualization (Van der Maaten & Hinton, 2008), as shown in Figure 7. Our BAIT exhibits notable unlearnability both against pretrained backbones and train-from-scratch backbones, which leads to the misclassification of clean test samples by the classifier, thereby providing effective protection of real semantics and demonstrating the applicability for practical applications.

F More Experiments

F.1 Performance with a Randomly Initialized Surrogate

To further evaluate the generalizability of the proposed method, we optimize BAIT with a randomly initialized surrogate model. We then evaluate its performance against the ImageNet-pretrained ResNet-18 model on three distinct datasets, including CIFAR-10, CIFAR-100, and SVHN. As shown in Table 12, our method outperforms baseline methods and exhibits strong capabilities in rendering data unlearnable, which further demonstrates its effectiveness against pretraining priors.

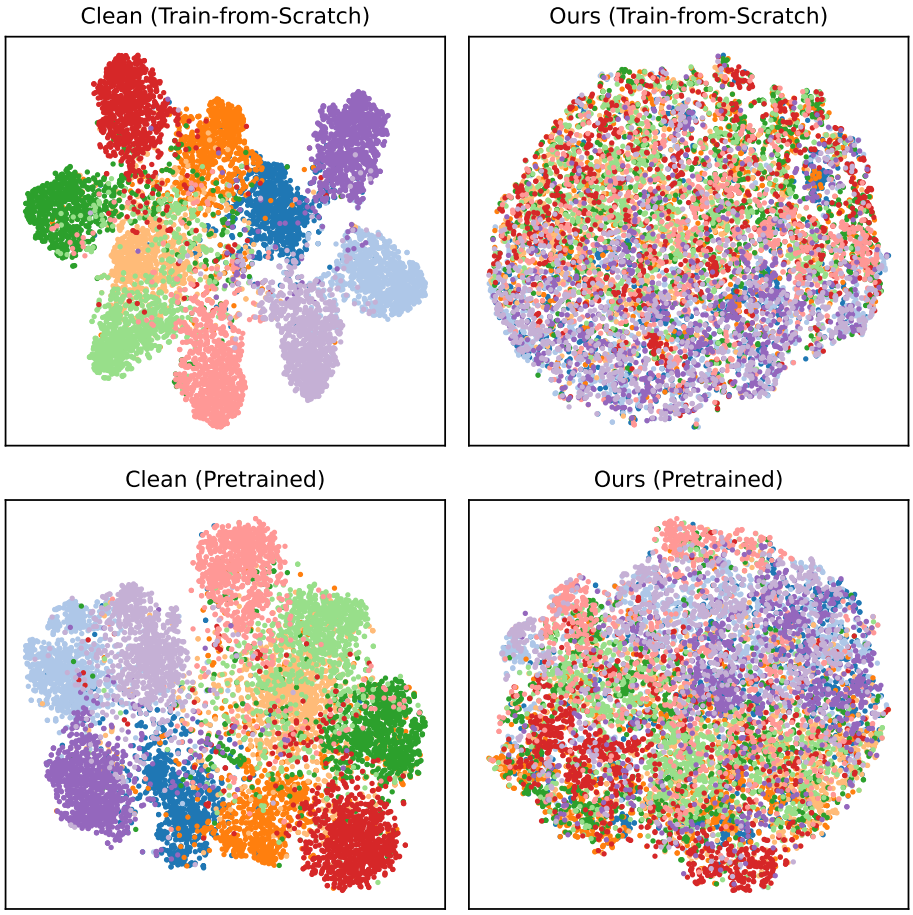


Figure 7: T-SNE visualization of classifier's last layer features, where classifiers are trained on the perturbed training set and tested on the clean test set. In the first row, models are trained on UEs in a train-from-scratch manner. In the second row, models are trained on UEs with pretraining priors.

Table 12: Test accuracy (%) ↓ comparison against ImageNet-pretrained backbones on CIFAR-10 and CIFAR-100, with a randomly initialized surrogate model for perturbation optimization,

Dataset	EMN	TUE	REM	LSP	GUE	14A	**Ours**
CIFAR-10	61.82	82.72	74.46	54.20	23.17	65.70	**15.97**
CIFAR-100	55.53	55.20	55.22	36.98	54.09	33.67	**22.83**
SVHN	37.55	41.15	76.45	38.91	39.68	81.47	**17.93**

F.2 Performance with ViT-based Models

To investigate whether larger models with more robust pretrained knowledge can exhibit superior resistance to UEs, we have added additional evaluations against more advanced or larger models, including Tiny-ViT (Wu et al., 2022), Swin Transformer Tiny (Liu et al., 2021), and ViT-B/16 (Dosovitskiy et al., 2021), on CIFAR-10 and CIFAR-100. We also evaluate the effectiveness of our method with these ViT-based models. As shown in Table 13, these models indeed exhibit stronger robustness when training with unlearnable data. For example, TUE achieves more than 90% test accuracy on CIFAR-10, and GUE obtains over 80% test accuracy on CIFAR-100. Even so, we observe that our method outperforms baselines clearly and still introduces unlearnability, which further demonstrates the efficacy of our proposed BAIT framework.

Table 13: Test accuracy (%) ↓ comparison against ImageNet-pretrained ViT models on CIFAR-10 and CIFAR-100 datasets.

Dataset	Backbone	EMN	TUE	REM	LSP	GUE	**Ours**
CIFAR-10	Tiny-ViT	76.00	95.53	87.83	79.66	25.04	**17.41**
	Swin Transformer Tiny	67.09	89.90	79.23	39.62	26.45	**15.67**
	ViT-b/16	89.53	97.22	92.66	90.06	44.79	**31.62**
CIFAR-100	Tiny-ViT	40.72	47.47	52.39	44.93	81.40	**23.64**
	Swin Transformer Tiny	39.12	40.89	53.85	53.99	86.85	**19.16**
	ViT-b/16	57.63	51.53	75.09	65.68	86.26	**23.30**

F.3 Cross-Task transferability

Table 14: Evaluation of cross-task transferability from classification to segmentation on Pascal VOC 2012.

Method	Pascal VOC 2012 Semantic (mIoU (%) ↓)				
	Aeroplane	Bicycle	Bird	Boat	Bottle
Clean	80.9	35.6	84.4	65.8	74.7
UnSeg	**30.5**	**16.4**	**58.6**	**19.1**	**40.4**
LSP	41.3	50.2	63.3	46.9	57.1
Ours	40.0	49.0	60.4	44.6	53.5

We conduct additional evaluations on the transferability of our method from classification to a different image segmentation task, with perturbations optimized from the CIFAR-10 classification dataset. We note that directly transferring perturbations may not be suitable since segmentation requires dense, pixel-wise predictions, whereas classification relies on single, image-level labels. This discrepancy is likely to impact the effectiveness of UEs. Following the evaluation setup of UnSeg (Sun et al., 2024), we conduct experiments on the Pascal VOC 2012 dataset (Everingham et al., 2010). Note that UnSeg is specifically designed for segmentation and is considered to be a very strong baseline. We also evaluate the cross-task transferability of LSP (Yu et al., 2022) for comparison. As shown in Table 14, we observe that although not explicitly designed for cross-task transferability, our method can reduce the mIoU for several classes compared to clean training, indicating that our method maintains certain unlearnability when transferred to other downstream tasks. Moreover, our method outperforms LSP across five classes, demonstrating superior cross-task transferability compared with other classification-oriented UE baselines. We acknowledge that there is still room for improvement in bridging the performance between our method and the segmentation-specialist method UnSeg. Given the broad potential of cross-task UEs, we consider this as an important future research direction.

F.4 Verification with different data protection ratios.

Although the standard UE setting applies perturbations to the entire training set, in practical scenarios, there may be only a portion of the data that is protected. To this end, following previous UE studies (Huang et al., 2021; Yu et al., 2022; Sadasivan et al., 2023), we also explore the impact of the data protection ratio on unlearnability. The experimental results are presented in Table 15. We observe that mixing clean samples with poisoned samples leads to a small performance increase compared to the clean training, indicating that models gain little semantic information from the UEs. Moreover, we observe that our method successfully reduces the test accuracy slightly compared to clean training at poison ratios of 10%, 20%, and 30%, and exhibits the smallest accuracy increase compared to baseline methods at other poison ratios, demonstrating its superior effectiveness.

Table 15: Test accuracy (%) ↓ under different poisoning ratios p against the ImageNet-pretrained ResNet-18 model on CIFAR-10.

Poison Ratio r	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Clean $(1-r)$	83.13	82.97	82.20	81.40	80.02	78.40	76.96	73.88	67.57	–
EMN	83.58	83.41	83.32	83.08	83.04	82.37	81.24	80.52	77.28	61.82
TUE	83.87	83.83	83.81	83.79	83.54	83.43	83.39	83.03	82.97	82.72
REM	84.18	83.79	83.63	83.05	82.69	82.90	81.96	81.09	79.81	74.46
LSP	83.56	83.51	82.53	82.49	81.48	80.29	79.68	77.97	74.91	54.20
GUE	83.94	83.58	83.48	82.48	81.79	81.14	79.88	78.18	74.50	23.17
Ours	**83.06**	**82.94**	**81.67**	**81.51**	**81.11**	**79.43**	**78.13**	**74.95**	**71.88**	**14.40**