Evading Data Provenance in Deep Neural Networks

Hongyu Zhu[1]* Sichu Liang[2]* Wenwen Wang[3] Zhuomeng Zhang[1] Fangqi Li[1] Shi-Lin Wang[1]†
[1]Shanghai Jiao Tong University [2]Southeast University [3]Carnegie Mellon University
Code: https://github.com/dbsxfz/EscapingDOV

Abstract

Modern over-parameterized deep models are highly data-dependent, with large scale general-purpose and domain-specific datasets serving as the bedrock for rapid advancements. However, many datasets are proprietary or contain sensitive information, making unrestricted model training problematic. In the open world where data thefts cannot be fully prevented, Dataset Ownership Verification (DOV) has emerged as a promising method to protect copyright by detecting unauthorized model training and tracing illicit activities. Due to its diversity and superior stealth, evading DOV is considered extremely challenging. However, this paper identifies that previous studies have relied on over-simplistic evasion attacks for evaluation, leading to a false sense of security. We introduce a unified evasion framework, in which a teacher model first learns from the copyright dataset and then transfers task-relevant yet identifier-independent domain knowledge to a surrogate student using an out-of-distribution (OOD) dataset as the intermediary. Leveraging Vision-Language Models and Large Language Models, we curate the most informative and reliable subsets from the OOD gallery set as the final transfer set, and propose selectively transferring task-oriented knowledge to achieve a better trade-off between generalization and evasion effectiveness. Experiments across diverse datasets covering eleven DOV methods demonstrate our approach simultaneously eliminates all copyright identifiers and significantly outperforms nine state-of-the-art evasion attacks in both generalization and effectiveness, with moderate computational overhead. As a proof of concept, we reveal key vulnerabilities in current DOV methods, highlighting the need for long-term development to enhance practicality.

1. Introduction

The success of modern deep learning hinges heavily on abundant datasets, spanning from large-scale vision and multi-modal resources such as ImageNet [9] and LAION-5B [52] to datasets tailored for vertical industries like face recognition [37] and computational pathology [71]. Curated meticulously through extensive human effort for collection, cleaning, and labeling, these datasets are protected as intellectual property and are available solely through licensed access. Some are released for academic purposes with restrictions on commercial exploitation, while others are proprietary assets, strictly prohibiting third-party model training. Moreover, personal data—such as photos shared on social platforms—is increasingly at risk of being scraped without consent [44]. Consequently, unauthorized model training on restricted data poses an escalating threat to both intellectual property and privacy rights.

Despite growing regulatory enforcement, like the General Data Protection Regulation (GDPR) [10], preventing unauthorized training remains a formidable challenge. Every stage of the data supply chain is vulnerable to attacks that funnel data into illicit markets [66], including insider threats [59], side-channel exploits [21], illegal mobile app access [80], and leakage from pretrained models [27, 70]. High-profile cases of data misuse in model training, such as the Cambridge Analytica scandal [69] and the Flickr Leakage [44], have intensified public scrutiny.

Consequently, preventing unauthorized model training at its source—the dataset itself—is crucial. Unlearnable Examples [23] introduce perturbations to prevent models from learning meaningful features. However, beyond their susceptibility to adaptive attacks [51], they entirely obstruct legitimate dataset use. As an alternative, post hoc verification for data provenance, known as **Dataset Ownership Verification** (DOV), has garnered substantial attention. Inspired by watermarks used to track duplication, remixing, or exploitation of multi-modal content [6], the *backdoor watermark* embeds carefully crafted "poison" samples selected by the dataset owner. Models trained on watermarked data exhibit predefined behaviors in response to backdoor triggers, enabling model-agnostic owner attribution [33, 63].

To mitigate the impact of backdoor on authorized training, recent techniques introduce *non-poisoning watermarks* that adjust prediction confidence without inducing misclassification, employing hypothesis testing for ownership trac-

*Equal contribution. †Corresponding author.

ing [50, 68, 79]. Non-intrusive *dataset fingerprint* have also emerged, allowing verification via elevated confidence on memorized training samples [35, 42]. DOV is widely regarded as an effective, if not the only, solution for tracking unauthorized training without disrupting legitimate usage [17, 32], with rapid advancements underway.

However, when deployed in the real world, the arms race between attackers and defenders forms an ongoing strategic game. Adversaries attempt to suppress verification behaviors to evade detection, while defenders continually strengthen verification mechanisms to reliably identify unauthorized training. Validating the robustness of DOV against evasion attempts is thus essential for practical deployment. Despite substantial progress in DOV, advances in evasion attacks remain underdeveloped, relying on simple regularization techniques and generic adaptations from poisoning defenses like fine-tuning [17, 32, 35, 79], which lack the versatility to counter diverse DOV techniques. Meanwhile, advanced DOV strategies, such as clean-label, invisible, or untargeted *backdoor watermarks* provide exceptional flexibility; *non-poisoning watermarks* and *dataset fingerprints*, which avoid misclassification or require no dataset modifications, remain nearly undetectable. It is widely agreed that designing a universal attack capable of evading all forms of verification is nearly impossible [42, 68]. This gap between the advancement of DOV and evasion techniques hinders reliable assessments.

In this paper we ask: is current DOV robust enough to preclude any evasion attempt? Our findings reveal that **prior evasion attacks are too weak, leading to a false sense of security**. Through analysis of verification behaviors, we identify shared characteristics: they are both *exclusive* and *subtle*, serving as unique markers that distinguish the protected dataset while remain incompatible with underlying semantic distribution of the main task.

Building on this insight, we propose the first universal evasion framework, **Escaping DOV**, which requires no additional clean in-distribution data or assumptions about the verification process. In Escaping DOV, a teacher model is initially trained directly on the copyright dataset to fully absorb domain knowledge. Then, an OOD transfer set, unrelated to the protected data, serves as an intermediary to extract *task-oriented* yet *identifier-free* knowledge to a surrogate student for deployment. Intuitively, exclusive and subtle identifiers, acting as side-channel signals orthogonal to the task distribution, make it improbable for OOD samples to trigger verification behaviors. Thus, only essential, identifier-invariant task knowledge is likely to transfer to the student model. To enhance Escaping DOV, we propose Transfer Set Curation, leveraging the unbiased knowledge embedded in large language models (LLMs) and vision-language models (VLMs) to retrieve the most *informative* and *reliable* samples from a general OOD gallery set. Additionally, we introduce Selective Knowledge Transfer to block suspicious verification behaviors, achieving better balance between generalization and evasion. The overall pipeline for Escaping DOV is shown in Figure 1.

Our contributions are summarized as follows:

1. We reveal that current DOV robustness evaluations are insufficiently rigorous. We introduce Escaping DOV, the first universal evasion framework that achieves simplicity, efficacy, and generalizability.
2. We propose Transfer Set Curation and Selective Knowledge Transfer to achieve balance between generalization and evasion with moderate overhead.
3. Experiments on diverse datasets and 11 DOV methods, alongside comparisons with 9 SOTA evasion methods, validate the efficacy of Escaping DOV, establishing it as a reliable benchmark for future DOV assessments.

2. Related Work

2.1. Dataset Ownership Verification

Backdoor attacks inject poisoned samples to associate predictions with a specific trigger [16], which induces misclassification whenever the trigger appears [15]. When controlled by copyright owners, backdoors can also act as identifiers for model ownership [1]. Inspired by this, the straightforward *backdoor watermark* embeds backdoor samples in the dataset. Models trained on this marked data produce a predefined label with high probability when evaluated on trigger samples [33]. The defender's *knowledge* of this secret prediction behavior serves as *proof* of ownership.

Early backdoors link triggers to target classes by manipulating the labels of poisoned samples. However, due to the obvious nature of *poisoned labels*, these samples are easily filtered through automatic data sanitization [73]. Clean-label backdoor watermark [63] utilizes advanced poisoning techniques [65] to enforce backdoor behavior without modifying labels. Later advancements further constrain modifications within an L_∞ norm ball [74], making such *clean-label, invisible* backdoor watermarks highly stealthy [73]. Recently, untargeted backdoor watermark [32] induces non-deterministic misclassification without a fixed target class, enhancing resilience against post-training defenses such as trigger synthesis and output distribution analysis.

Advanced backdoor watermark provides a stealthy and resilient verification mechanism. However, it inevitably embeds harmful shortcuts [76], problematic when authorized training is essential [17]. *Non-poisoning watermark* subtly embeds watermark features into the dataset without inducing misclassification. Verification is performed by measuring loss differences between watermarked and clean samples via hypothesis testing. Various mechanisms have been developed to generate stealthy watermarks: Radioactive data [50] and ML Auditor [24] optimize

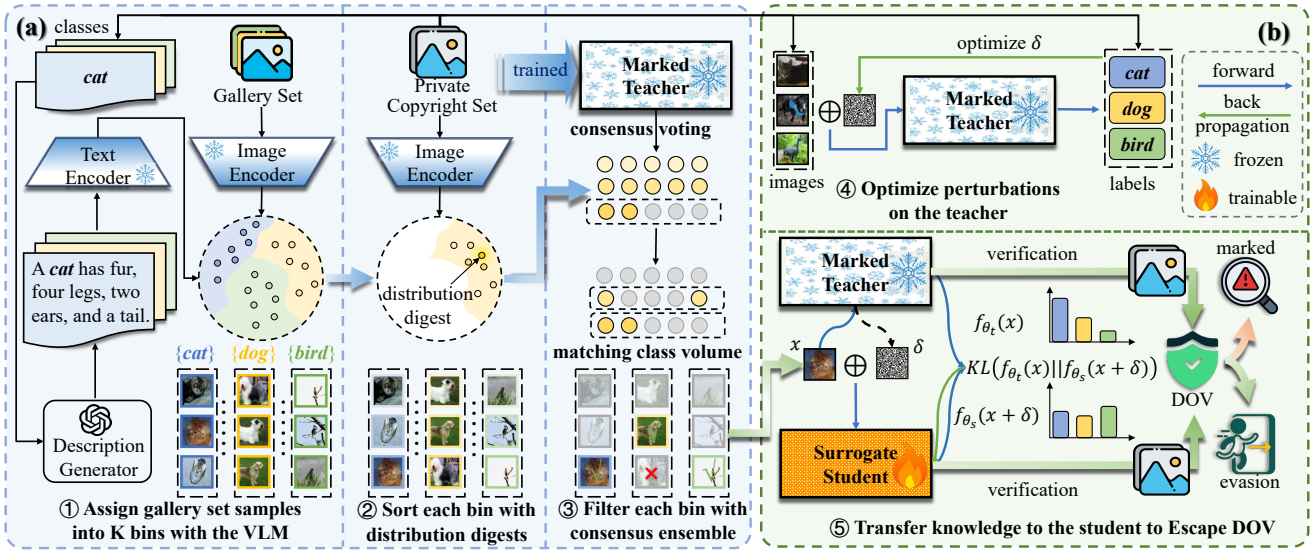


Figure 1. **Overall Pipeline of Escaping DOV. (a) Transfer Set Curation** (§3.2): Samples from the OOD gallery set are clustered via a VLM, ranked by distribution proximity, and filtered through consensus ensemble to curate an optimal transfer set. **(b) Selective Knowledge Transfer** (§3.3): Promoting invariance to teacher's worst perturbations, facilitating task-oriented yet identifier-free knowledge transfer.

L_∞-perturbations, Anti-Neuron Watermarking [79] applies color transformations in hue space, Domain Watermark [17] converts samples into a hardly-generalizable domain, while Isotope [68] blends external features into the dataset.

Additionally, *dataset fingerprint* utilizes intrinsic features for verification without modifying the dataset. Dataset inference [42] analyzes the margins between training and external samples relative to the decision boundary, while MeFA [35] aggregates membership inference across multiple samples. Although dataset fingerprints are non-intrusive and support post hoc protection for published datasets, they are susceptible to high false positives [61, 62].

2.2. Evasion Attacks against Data Provenance

Despite rapid progress in DOV, current evasion strategies rely on basic regularization and backdoor defenses like fine-tuning [17, 32, 36]. **As they barely evade any DOV methods, we present their results in Appendix C.1 and use SOTA backdoor defenses as baselines**, which are categorized by stage of application: (1) *Data sanitization* removes suspicious samples *pre-training* [73]; (2) *Robust learning* mitigates poisoning effects *during training* [4, 31]; (3) *Backdoor unlearning* synthesizes triggers *post-training* to neutralize the model [72]; and (4) *Input preprocessing* detects or purifies triggers *during testing* [56]. Together, these defenses span the model lifecycle; yet advanced invisible, clean-label, or untargeted backdoor watermarks can bypass most of them [17, 32]. Further, non-poisoning watermarks and non-intrusive fingerprints render backdoor defenses ineffective, making universal evasion extremely challenging. This work is the first to establish a unified evasion perspective, opening new directions for robust evaluation of DOV.

We discuss the contemporaneous works [77] and [55] that emerged at the time of acceptance in Appendix E.

3. Escaping DOV

3.1. Overview

In K-class image classification, an adversary gains unauthorized access to the copyright dataset $\mathcal{D}$ to train a model f_θ, which maps images $x \in \{0, 1, \ldots, 255\}^{C\times W\times H}$ to predictions $y \in \mathbb{R}^K$. For data provenance, the defender employs a DOV method $\text{Verif}(f_\theta, \mathcal{D}, \mathcal{M})$, inspecting the output of the suspicious model f_θ on specific queries. This process yields a verification metric indicating the likelihood that f_θ was trained on $\mathcal{D}$, such as the Attack Success Rate for backdoor watermarks or the confidence for rejecting the null hypothesis in non-poisoning DOV. Here, $\mathcal{M}$ represents external materials (e.g., watermark trigger patterns) controlled by the defender to uniquely identify the dataset.

DOV methods are flexible and diverse. Backdoor watermark achieves superior stealth through invisible, clean-label, or untargeted poison techniques. Non-poisoning watermarks avoid misclassification, while dataset fingerprints require no dataset modifications. Hence, direct sanitization or suppression of all potential verification behaviors becomes a formidable challenge; most evasion attempts that simply adapt poison defenses have consistently failed [17, 32, 42, 68, 79]. However, we identify a common trait among all DOV strategies: verification behaviors are both *exclusive* and *subtle*. On one hand, these behaviors must *exclusively* belong to the copyright dataset to prevent false attribution to models trained on external datasets. On the other hand, they must be *subtle* enough to remain concealed

during data sanitization, ensuring they are rarely triggered by regular samples to preserve the generalization of models trained for authorized use.

Accordingly, we propose a universal evasion strategy: first, a teacher model f_{θ_t} is trained on the copyright dataset $\mathcal{D}$ to absorb task-domain knowledge, inevitably marked by DOV. Next, an unrelated OOD transfer set $\mathcal{T}$ serves as the medium to extract task-oriented yet identifier-invariant knowledge into a surrogate student f_{θ_s}. Intuitively, the *exclusive* and *subtle* watermark triggers are absent in the natural samples of $\mathcal{T}$, preventing the activation and transmission of verification behaviors to the surrogate student. For fingerprints based on exaggerated confidence, the student only indirectly inherits knowledge from the marked teacher via the unrelated OOD data, significantly reducing excessive memorization associated with $\mathcal{D}$.

The overall objective of Escaping DOV can be formalized as simultaneously minimizing the generalization error on the task distribution and the verification metric:

$$\min_{\theta_s} \mathbb{E}_{(x,y)\in P(x,y)}[\mathcal{L}(f_{\theta_s}(x), y)] + \alpha \cdot \text{Verif}(f_{\theta_s}, \mathcal{D}, \mathcal{M}) \quad (1)$$

Here, f_{θ_s} represents the surrogate student ultimately deployed, $P(x, y)$ the underlying distribution of samples $x, y \in \mathcal{D}$, and $\mathcal{L}$ the loss function measuring prediction discrepancies. The tuning factor α balances the pursuit of generalization and evasion for f_{θ_s}. However, $P(x, y)$ and the process of $\text{Verif}(f_{\theta_s}, \mathcal{D}, \mathcal{M})$ are agnostic to the adversary and intractable during optimization. Therefore, we propose two modules to better achieve and balance these objectives: (1) For *generalization*, we curate the most informative and reliable transfer set from a large-scale OOD gallery set to support high-performance and identifier-invariant knowledge extraction; (2) For *evasion*, to prevent the transmission of suspicious verification knowledge via dark knowledge in soft labels, we propose selective knowledge transfer to filter out such knowledge and better trade off between the two objectives. Next, we elaborate on these two components.

3.2. Transfer Set Curation

Our objective is to curate an optimal subset $\mathcal{T}$ from an OOD gallery set $\mathcal{G}$ (e.g., ImageNet [9] or DataComp [12]), selecting samples that are both *informative* and *reliable* to facilitate effective knowledge transfer. While this task may resemble classical core-set selection [5, 53], it presents significantly greater challenge: (1) Our algorithm operates in a fully OOD context, where $\mathcal{G}$ and $\mathcal{D}$ do not necessarily overlap in classes, and their distributions could vary substantially; (2) $\mathcal{D}$ could include identifiers manipulated by the defender (e.g., trigger samples), and the teacher f_{θ_t} trained directly on $\mathcal{D}$ is already marked. Consequently, selection based on the unreliable $\mathcal{D}$ and f_{θ_t} inadvertently introduce samples from $\mathcal{G}$ that activate verification behaviors, risking the induction of these behaviors in the surrogate student f_{θ_s}.

In this paper, we leverage a vision-language model (VLM) with unbiased pretrained knowledge and robust zero-shot capabilities (e.g., CLIP [49]) to facilitate reliable selection alongside the marked teacher f_{θ_t}. CLIP aligns visual and textual features in a shared embedding space, where the image encoder $\phi_I(\cdot)$ maps an image x into $\phi_I(x)$, and text templates with category names (e.g., "a photo of {class name}") $c_1, \ldots, c_K \in \mathcal{D}$ are mapped by the text encoder $\phi_t(\cdot)$ to $\phi_t(c_1), \ldots, \phi_t(c_K)$. By calculating the cosine similarity between $\phi_I(x)$ and $\phi_t(c_i)$, each sample $x \in \mathcal{G}$ can be uniquely assigned to a category t in $\mathcal{D}$.

However, class-name-only text templates lack foreground semantics [48], limiting the ability to distinguish target categories. While few-shot tuning allows the VLM to rapidly adapt to tasks within $\mathcal{D}$ [26], it risks being compromised by verification behaviors embedded in $\mathcal{D}$ [3]. Inspired by advances in zero-shot prompt learning [48], we adjust the VLM with unbiased category semantics from a large language model (LLM). The LLM generates a description set Desc_{c_i} for each category c_i in $\mathcal{D}$, averaging descriptions in the feature space to create representation prototypes for each category, enhancing the zero-shot capabilities of the VLM. Consequently, each $x \in \mathcal{G}$ is assigned to class t in a zero-shot fashion as follows in Equation 2:

$$t = \arg\max_{c_i} \frac{1}{|\text{Desc}_{c_i}|} \sum_{d \in \text{Desc}_{c_i}} (\text{sim}(\phi_I(x), \phi_I(d))) \quad (2)$$

where sim denotes cosine similarity shown in Equation 3:

$$\text{sim}(\phi_I(x), \phi_I(d)) = \frac{\phi_I(x) \cdot \phi_I(d)}{|\phi_I(x)| \cdot |\phi_I(d)|} \quad (3)$$

Although the VLM assigns all gallery set samples to categories in $\mathcal{D}$, it lacks a criterion for the most *informative* samples. To leverage the distribution from $\mathcal{D}$ while avoiding biases from potential verification behaviors, we project all samples in $\mathcal{D}$ into the feature space with $\phi_I(\cdot)$ and calculate the density centroids Cent_t for each class $t \in \{1, \ldots, K\}$. These centroids, acting as **distribution digests**, encapsulate the conditional distribution of $\mathcal{D}$. Samples closer to Cent_t in feature space are prioritized for selection, a robust and straightforward criterion that is theoretically resistant to manipulation by a small proportion of outlier samples [25].

We then establish a consensus-based ensemble for final transfer set curation: (1) The gallery set $\mathcal{G}$ is divided by the LLM-enhanced VLM into K bins corresponding to classes in $\mathcal{D}$. Within each bin $\mathcal{G}_t$, (2) samples are ordered by proximity to distribution digest Cent_t; (3) Only samples classified by the teacher f_{θ_t} as t that are closest to Cent_t are selected, until the selected samples match the class count $|\mathcal{D}_t|$. Illustrated in Figure 1, this process ensures all predictions on the transfer set $\mathcal{T}$ by f_{θ_t} are consistent with VLM, minimizing the risk of triggering verification behaviors while closely approximating the distribution in $\mathcal{D}$.

The Feature Bank. The curation process involves projecting all gallery set samples into the feature space, which is the most time-consuming step. However, when the gallery set is potentially reused, this projection only needs to be performed once. The resulting features and indexes are saved in a **feature bank** for efficient reuse.

3.3. Selective Knowledge Transfer

The transfer set $\mathcal{T}$ serves as an intermediary for extraction of task-oriented and identifier-invariant knowledge. This process employs the standard knowledge distillation framework [20] with the following optimization objective:

$$\theta_s^* = \arg\min_{\theta_s} \mathcal{L}\left(\frac{f_{\theta_s}(x)}{\tau}, \frac{f_{\theta_t}(x)}{\tau}\right), \quad x \in \mathcal{T} \tag{4}$$

Here, $\mathcal{L}$ measures prediction discrepancies, such as the Kullback-Leibler divergence, modulated by the temperature τ. This process effectively transfers knowledge despite potential misalignments between the transfer set $\mathcal{T}$ and the training distribution [46]. However, the soft labels $f_{\theta_t}(x)$ contain dark knowledge [11], which inadvertently transmit verification behaviors. Thus, we introduce Selective Knowledge Transfer (SKT) to filter out suspicious verification knowledge during distillation.

The *exclusive* and *subtle* verification behaviors reflect unique biases inherent to the marked dataset $\mathcal{D}$, distinguishing it within the task distribution. Preventing the student f_{θ_s} from overfitting to biases limits DOV's ability to infer connections between f_{θ_s} and $\mathcal{D}$ through trigger-specific outputs or excessive confidence. Drawing on insights from shortcut learning [14], we mitigate predictive shortcuts during the student's learning process, reducing reliance on spurious features. Since dataset biases typically induce universal shortcut behaviors, a straightforward solution is to apply universal adversarial training (UAT) on the student:

$$\theta_s^* = \arg\min_{\theta_s} \max_{\delta} \frac{1}{n} \sum_{x,y \in \mathcal{T}} \mathcal{L}(f_{\theta_s}(x+\delta), y) \tag{5}$$

UAT effectively suppresses atypical predictive behaviors in f_{θ_s} induced by perturbations δ. However, the min-max problem involves a notoriously hard bi-level optimization, requiring iterative solutions within each iteration. Furthermore, the absence of ground-truth labels in the OOD transfer set $\mathcal{T}$ necessitates predictions from the marked teacher as labels, introducing significant calibration errors.

Given the objective to mitigate biases specific to the copyrighted dataset $\mathcal{D}$, and with all verification behaviors encapsulated in the teacher model, we propose using only the teacher model f_{θ_t} to generate perturbations on $\mathcal{D}$ as a practical approximation for the inner max-problem:

$$\delta^* = \arg\max_{\delta} \frac{1}{n} \sum_{x,y \in \mathcal{D}} \mathcal{L}(f_{\theta_t}(x+\delta), y) \tag{6}$$

This strategy decouples the bi-level optimization, rendering it asynchronously solvable. By pre-generating a **perturbation pool** $\{\delta_i\}_{i=1}^n$ in an offline manner with the teacher f_{θ_t} and reusing it in the outer min-problem during distillation, computational overhead is significantly reduced.

To solve Equation 6, the perturbation δ is typically constrained by an L_p norm to preserve core image semantics [40]. However, without prior knowledge of the specific form of δ associated with verification behaviors (e.g., watermark triggers), choosing an inappropriate norm undermine the efficacy of UAT in suppressing spurious features [64]. Thus, we seek to diversify the teacher-generated perturbations beyond a single norm constraint.

We solve the max-problem using mini-batch stochastic gradient descent [54], applying L_2 constraints at each batch update to preserve gradient directions. After each iteration, we project the perturbation onto the norm constraint (L_0, L_2, or L_∞) that maximizes the loss. This steepest descent projection [41] adaptively seeks the optimal norm constraint, facilitating the generation of perturbations closely resembling those employed by DOV. Moreover, extreme projections onto L_0 or L_∞ norms approximate a convex hull spanning multiple norm constraints, broadening the range of potential perturbations [7].

Ideally, perturbation should preserve core image semantics without being limited to specific norms [57]. Generative models can achieve this by producing unrestricted adversarial perturbations [76], though training such models on task-specific distributions is resource-intensive. Therefore, we propose a lightweight approach using image corruptions instead of generative models to produce unrestricted perturbations. Starting with the 15 corruptions defined in ImageNet-C [19], we employ a genetic algorithm [8, 43] to select the combination and sequence of corruptions that solves Equation 6, forming a corruption chain:

$$\max \mathcal{L}(f_{\theta_t}(\text{Corrupt}_{1:N}(x)), y), \quad x, y \in \mathcal{D} \tag{7}$$

The perturbations and corruption chain generated with the teacher f_{θ_t} on $\mathcal{D}$ are collectively referred to as $A(\cdot)$, leading to the final objective for Selective Knowledge Transfer:

$$\begin{aligned} \arg\min_{\theta_s} \mathcal{L}&\left(\frac{f_{\theta_s}(x)}{\tau}, \frac{f_{\theta_t}(x)}{\tau}\right) \\ &+ \beta \cdot \mathcal{L}\left(\frac{f_{\theta_s}(A(x))}{\tau}, \frac{f_{\theta_t}(x)}{\tau}\right), \quad x \in \mathcal{T} \end{aligned} \tag{8}$$

The second term promotes the student's invariance to the worst-case perturbations from the teacher while encouraging predictive divergence to reduce bias toward the teacher. The tuning factor β balances generalization with evasion efficacy, applied as a sampling probability to selectively introduce perturbations or corruption chain to a subset of samples, adding minimal overhead.

Table 1. Escaping DOV on CIFAR-10 and Tiny ImageNet. Red indicates detection by DOV, while green indicates successful evasion.

DOV	CIFAR-10				Tiny-Imagenet			
	Vanilla		Evasion		Vanilla		Evasion	
Poisoning DOV	ACC	VSR	ACC(↑)	VSR(↓)	ACC	VSR	ACC(↑)	VSR(↓)
Badnets	94.44 ± 0.74	100.00 ± 0.00	93.46 ± 0.20	1.36 ± 0.07	69.05 ± 0.25	98.98 ± 0.11	66.48 ± 0.14	1.48 ± 0.11
UBW	93.98 ± 0.25	95.54 ± 0.84	93.41 ± 0.19	1.74 ± 0.06	66.37 ± 0.12	75.56 ± 0.50	65.34 ± 0.08	4.60 ± 0.11
Label-Consistent	94.74 ± 0.54	96.18 ± 2.70	93.19 ± 0.14	3.74 ± 0.40	69.07 ± 0.24	34.99 ± 1.53	66.46 ± 0.10	3.51 ± 0.22
Narcissus	94.76 ± 0.54	87.34 ± 1.82	94.37 ± 0.18	4.59 ± 0.64	68.99 ± 0.14	57.49 ± 0.78	66.39 ± 0.18	12.33 ± 6.45
Non-Poisoning DOV	ACC	p-value	ACC(↑)	p-value(↑)	ACC	p-value	ACC(↑)	p-value(↑)
Radioactive Data	94.96 ± 0.47	3.03e-03	94.07 ± 0.18	9.45e-01	68.84 ± 0.30	5.33e-03	66.64 ± 0.24	4.94e-01
ANW	94.66 ± 0.31	1.37e-09	93.91 ± 0.28	1.00e+00	68.75 ± 0.30	5.57e-28	66.44 ± 0.13	1.00e+00
Domain Watermark	94.38 ± 0.53	1.67e-22	93.90 ± 0.13	1.00e+00	68.55 ± 0.39	3.91e-16	66.38 ± 0.16	1.00e+00
Isotope	94.75 ± 0.28	2.87e-03	93.99 ± 0.49	2.84e-01	68.97 ± 0.37	1.63e-03	66.76 ± 0.22	1.66e-01
ML Auditor	94.22 ± 0.17	7.47e-05	93.45 ± 0.21	7.83e-01	68.41 ± 0.28	1.55e-06	65.35 ± 0.16	1.00e+00
Dataset Inference	94.83 ± 0.47	1.87e-03	93.97 ± 0.49	4.76e-01	69.15 ± 0.34	9.27e-14	66.06 ± 0.18	4.47e-01
MeFA	94.83 ± 0.47	2.62e-14	93.93 ± 0.55	1.00e+00	69.15 ± 0.34	6.17e-27	66.22 ± 0.20	9.99e-01

4. Experiments

4.1. Experimental Setup

Following established DOV methods [17, 33, 42, 79], we adopt ResNet-18 [18] for both teacher and students. For transfer set curation, we utilize MobileCLIP [67] and GPT-4o mini [45] as lightweight VLM and LLM, respectively.

We assess backdoor watermarks with Verification Success Rate (VSR)-the probability of a trigger-induced misclassification to the target class. **A VSR above 30% confirms successful verification** [76], indicating unauthorized training. Non-poisoning watermarks and fingerprints, collectively termed non-poisoning DOV, are evaluated via the p-value from a one-tailed T-test. The p-value represents the confidence level that the model was **not** trained on the copyright dataset; **lower p-values indicate successful verification**. The common threshold for p-values is set at 0.01 [42], equating to a 99% confidence level of unauthorized training. Experiments were repeated 3 times, reporting mean ± std for ACC and VSR, and harmonic mean for p-values.

4.2. Escaping Advanced DOV Methods

We evaluate Escaping DOV on natural image datasets commonly used as DOV benchmarks, including CIFAR-10 [28] and the 200-class Tiny ImageNet [29]. **We further consider six datasets with significant distribution shifts from the gallery set** (e.g., facial recognition and medical diagnosis), with a summary of results in Section 4.4.5 and detailed findings in Appendix C.3. For the gallery set, we use LSVRC-2012 [9]. Given overlapping samples between Tiny ImageNet and LSVRC-2012, we creat a challenging setting by *removing all overlapping classes from LSVRC-2012 to produce a purely OOD gallery set*. The targeted DOV approaches cover advanced algorithms across all categories: (1) backdoor watermarks, including BadNets [16], UBW [32], Label-Consistent [65], and Narcissus [74], encompassing clean-label, invisible, and untargeted poisoning techniques; (2) non-poisoning watermarks, including Radioactive Data [50], ANW [79], Domain Watermark [17], Isotope [68], and ML Auditor [24]; and (3) fingerprints, including Dataset Inference [42] and MeFA [35].

Table 1 presents the generalization and verification metrics for the models before and after evasion attacks (the teacher and surrogate student in Escaping DOV). All DOV methods successfully identify the *vanilla* teacher trained directly on the copyright dataset. However, none meet their verification thresholds (VSR $> 30\%$ or p-value < 0.01) on the student produced by Escaping DOV, where VSR is near random guesses and p-values exceed 0.1 in most cases. Moreover, Escaping DOV incurs minimal impact on generalization: test accuracy on CIFAR-10 decreases by less than 1%, while on the challenging Tiny ImageNet set, with no overlapping gallery classes, accuracy drops by only 2-3%. Thus, Escaping DOV achieves a robust balance between universal evasion and preserved generalization.

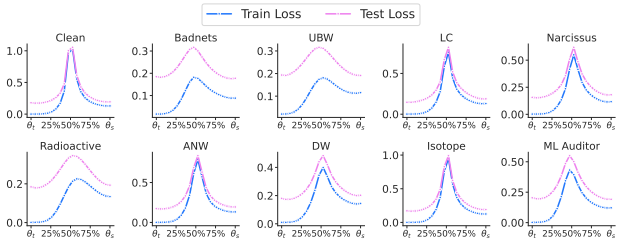


Figure 2. Loss Barrier between Teacher and Student Parameters.

We further observe that Escaping DOV encourages the student to diverge from the teacher's predictive behavior, forming its own mechanism. In Figure 2, we linearly interpolate between teacher and student parameters, computing training and test losses at each point. The results reveal two key findings: (1) A loss barrier appears during interpolation from teacher θ_t to student θ_s. This lack of linear connectivity suggests that the surrogate student develops a distinct prediction *mechanism*, bypassing the spurious features central to the teacher's verification behavior [39]. (2)

Table 2. Comparison of Escaping DOV with Evasion Attacks. Red indicates detection by DOV, green indicates successful evasion.

Method	Badnets		Narcissus		Isotope		Dataset Inference	
	ACC(↑)	VSR(↓)	ACC(↑)	VSR(↓)	ACC(↑)	p-value(↑)	ACC(↑)	p-value(↑)
Fine-pruning	86.21 ± 0.16	99.37 ± 0.27	86.58 ± 0.15	43.75 ± 14.41	86.78 ± 0.12	0.0532	86.88 ± 0.48	0.0095
Meta-Sift	88.68 ± 0.45	99.70 ± 0.32	88.64 ± 0.62	49.40 ± 8.26	88.56 ± 0.47	0.0264	88.07 ± 0.21	0.0963
Differential Privacy	90.49 ± 1.20	76.70 ± 4.62	87.70 ± 0.68	72.77 ± 3.91	88.69 ± 2.61	0.0678	89.64 ± 4.86	0.0440
I-BAU	89.09 ± 0.25	14.80 ± 0.46	88.76 ± 1.94	72.74 ± 2.17	87.07 ± 2.59	0.1552	89.93 ± 0.68	0.0355
ZIP	82.44 ± 0.10	79.88 ± 0.39	83.51 ± 0.37	53.21 ± 4.47	82.97 ± 0.42	0.0098	83.21 ± 0.90	0.1050
NAD	89.86 ± 0.37	3.11 ± 0.51	91.04 ± 0.36	75.41 ± 5.12	91.77 ± 0.45	0.0961	92.12 ± 0.34	0.0014
BCU	92.75 ± 0.05	1.56 ± 0.25	92.33 ± 0.11	61.72 ± 6.43	92.77 ± 0.36	0.1622	93.09 ± 0.18	0.0773
ABD	84.94 ± 3.04	7.07 ± 3.15	85.42 ± 2.10	22.10 ± 1.15	86.29 ± 3.46	0.0066	84.87 ± 3.02	0.3696
IPRemoval	83.96 ± 0.23	8.09 ± 0.50	85.06 ± 0.11	55.97 ± 2.14	84.64 ± 0.06	0.1916	85.28 ± 0.05	0.1322
Escaping DOV (Ours)	**93.46 ± 0.20**	**1.36 ± 0.07**	**94.37 ± 0.18**	**4.59 ± 0.64**	**93.99 ± 0.49**	**0.2845**	**93.97 ± 0.49**	**0.4759**

For the *teacher*, test loss is over 10^3 times the training loss, allowing fingerprints to easily infer dataset association. In contrast, the *student's* test loss is typically under twice the training loss, falling within the variance across subsets and making it far less distinguishable. Thus, Escaping DOV significantly suppresses both watermarks and fingerprints.

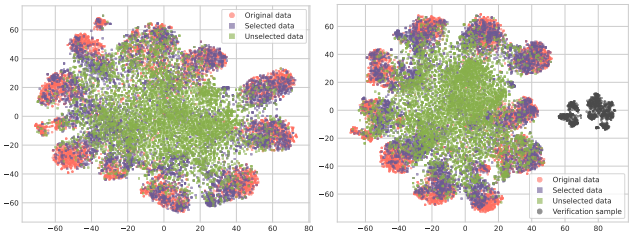
4.3. Comparison with SOTA Evasion Attacks

We compare with SOTA evasion attacks on CIFAR-10 across the seminal Badnets watermark and three most resistant DOVs from each category—Narcissus, Isotope, and Dataset Inference (see Table 1). Results for all DOV methods are provided in Appendix C.2. Evasion attacks include Fine-Pruning [36], stronger combination of fine-tuning and pruning widely used in DOV evaluations, along with SOTA backdoor defenses across four stages: data sanitization with Meta-Sift [73], robust training with Differential Privacy [4], backdoor unlearning with I-BAU [72], and input purification with ZIP [56]. We also compare with methods that use knowledge distillation to mitigate backdoors, such as NAD [31], BCU [47] and ABD [22], as well as IPRemoval [78], which aims to eliminate model watermarks. Results in Table 2 and Appendix C.2 show that Escaping DOV consistently surpasses all comparison methods in both generalization and evasion, and is the only approach that fully evades all DOV methods. For instance, only ABD and Escaping DOV reduce VSR below 40% on the clean-label and invisible Narcissus watermark; however, ABD fails to evade Isotope, while Escaping DOV achieves nearly 10% higher accuracy with superior evasion performance. We provide a rigorous analysis in Appendix D, explaining *why SOTA evasion attacks, particularly those relying on distillation, are less effective than our Escaping DOV*.

4.4. Ablations and Discussions

4.4.1. Transfer Set Curation

We begin with a *qualitative analysis* of the transfer set. Figure 3a visualizes the feature space for CIFAR-10 samples as well as selected and unselected ones from the transfer set, on a clean ResNet-18. The selected samples are notably closer to original CIFAR-10 samples. Figure 3b illustrates the feature space of a Badnets-marked model, where trigger samples form distinct clusters with almost no transfer set samples nearby. The *exclusive* and *subtle* verification behaviors are rarely activated by OOD samples, and the curation process effectively excludes suspicious samples.



(a) Clean (b) Badnets

Figure 3. Sample Distribution in the Feature Space.

Quantitative analysis follows, using Optimal Transport Dataset Distance (OTDD) [2] to assess the distance from the target dataset to each proxy distribution, as shown in Table 3. The curated transfer set is significantly closer to the target distribution than random sampling. Further, we compare the curated transfer set with random and unrelative samples on CIFAR-10 with Narcissus. Table 4 shows that transfer set curation not only improves generalization but also suppresses DOV by selecting more reliable samples.

Table 3. Optimal Transport Distances Between Datasets.

Distance	self	LSVRC(**curation**)	LSVRC(random)	GTSRB [58]
CIFAR-10	0.771	0.937	1.116	1.129
Tiny-Imagenet	0.855	0.988	0.993	1.066

Table 4. Escaping DOV across different Transfer Sets.

Transfer Set	GTSRB [58]	LSVRC(random)	LSVRC(**curation**)
ACC(↑)	66.80 ± 0.41	91.63 ± 0.49	**94.37 ± 0.18**
VSR(↓)	8.96 ± 1.63	7.87 ± 0.52	**4.59 ± 0.64**

4.4.2. Selective Knowledge Transfer

Figure 4 illustrates the impact of Selective Knowledge Transfer (SKT) on the most resistant Narcissus watermark across distillation temperatures τ. As τ increases, omitting SKT causes a substantial increase in VSR, while SKT consistently suppresses verification behavior across all temper-

atures. Additionally, VSR escalates sharply at higher temperatures, while test accuracy remains nearly unchanged. We hypothesis that task and verification knowledge possess distinct "boiling points", naturally separating at lower temperatures. Thus, we adopt $\tau = 1$ as a reliable default.

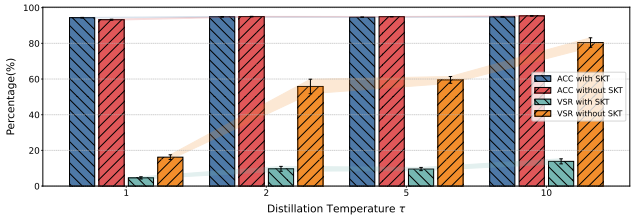


Figure 4. Impact of SKT and Distillation Temperature.

We further compare the offline perturbation pool in SKT with direct UAT on the student. While UAT lowers VSR to 2.07%, it incurs a 3.54% accuracy reduction and more than triples computational cost. As SKT balances evasion and generalization, it serves as a lightweight approximation. Detailed time complexity is analysed in Appendix C.8.

4.4.3. Escaping Adaptive Defenses

We assume a *knowledgeable* defender employing adaptive verification, specifically a semantic backdoor [60] and an anti-distillation backdoor [13]. For the semantic backdoor, we assume the defender knows the gallery set and selects 500 "tench" samples from LSVRC-2012, relabeling them as "plane" and inserting them into CIFAR-10 to test their classification probability as "plane" (VSR). Since the teacher memorizes this verification behavior, using it for sample selection allows triggers enter the transfer set. For instance, DFND [13], an entropy-based selection method, produces a student with VSR of 70.2% by including 254 "tench" samples in the transfer set. In contrast, our curated transfer set contains fewer than 10 "tench" samples, reducing VSR to only 17.4%. Even if the defender knows the gallery set, Transfer Set Curation still selects reliable samples. We also evaluate the entropy of teacher annotations: DFND's entropy is 2.90, while ours reaches 3.32, near the theoretical optimum $\log_2(10)$, indicating better class balance.

Table 5. Escaping DOV against the Adaptive ADB Watermark.

μ	Vanilla		Evasion		Independent	
	ACC	VSR	ACC	VSR	ACC	VSR
0.1	91.59%	100.00%	90.14%	94.42%	95.13%	92.89%
1	90.90%	100.00%	89.23%	78.48%	95.13%	71.74%
10	91.26%	12.88%	88.66%	14.21%	95.13%	10.01%

The anti-distillation backdoor (ADB) [13] is designed to transfer to student during distillation. As we employ distillation for knowledge transfer, we test ADB under a **hypothetical scenario** where **the defender controls the teacher's training process to optimize the trigger**. Table 5 reports the VSR across teacher, student, and unmarked models. A smaller μ allows larger perturbations, turning the trigger into an adversarial perturbation that induces **false positives** on unmarked models. Despite the defender's **unrealistic** control over the teacher, students' VSR remains closer to that of an unmarked model. As μ increases, ADB becomes ineffective. Thus, even under defender-controlled teacher training, Escaping DOV remains a potent threat.

4.4.4. Enhanced Evasion Tactics

Escaping DOV involves only post-training adjustments and operates orthogonally to pre-training, in-training and test-time strategies, all of which can collectively enhance evasion. As shown in Table 6, replacing data augmentation with mixup [75] during teacher training—an approach that mitigates backdoors and overfitting—further boosts evasion efficacy in conjunction with Escaping DOV on CIFAR-10.

Table 6. Mixup Teacher Training for Escaping DOV.

DOV	**Vanilla**		**Evasion**	
Poisoning DOV	ACC	VSR	ACC(↑)	VSR(↓)
Narcissus	94.04%	72.47%	93.03%	3.66%
Non-poisoning DOV	ACC	p-value	ACC(↑)	p-value(↑)
Isotope	94.03%	0.2964	92.71%	0.7013
Dataset Inference	94.60%	0.0937	92.78%	0.5196

4.4.5. Case Study on Domain-Specific Datasets

Datasets in vertical industries often require stronger protection than natural images due to collection challenges and privacy concerns. We conduct a case study on RAFDB (facial emotion) [30] and OrganCMNIST (medical imaging) [71] in Table 7, with detailed results in Appendix C.3. Escaping DOV maintains generalization and efficacy despite unique distribution, while a larger domain-related gallery and VLM could further improve performance [34, 38].

Table 7. Escaping DOV on Domain-specific Datasets.

Settings		**Vanilla**		**Evasion**	
DOV	**DataSet**	ACC	VSR	ACC(↑)	VSR(↓)
Badnets	RAFDB	85.21 ± 0.33	99.95 ± 0.02	80.53 ± 0.42	2.19 ± 0.06
	OrganC	95.32 ± 0.06	99.99 ± 0.01	90.57 ± 0.70	0.83 ± 0.15
DOV	**DataSet**	ACC	p-value	ACC(↑)	p-value(↑)
Isotope	RAFDB	82.46 ± 0.59	1.50e-06	80.37 ± 0.13	1.05e-01
	OrganC	94.63 ± 0.19	2.65e-03	88.96 ± 0.56	5.04e-01

5. Conclusion

In this paper, we reveal that Data Ownership Verification, widely used to prevent unauthorized model training, is vulnerable to strong evasion attacks. We introduce Escaping DOV, which leverages transfer set curation and selective knowledge transfer to achieve superior generalization and universal evasion capabilities. We propose that Escaping DOV serve as a evaluation framework to help establish truly reliable DOV methods, with plans for extensive cross-modal experiments to explore broader societal benefits.

Acknowledgments

The work was supported by the National Natural Science Foundation of China under Grant 62271307 and 61771310.

References

[1] Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX security symposium (USENIX Security 18)*, pages 1615–1631, 2018. 2

[2] David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020. 7

[3] Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24239–24250, 2024. 4

[4] Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. *Advances in Neural Information Processing Systems*, 36, 2023. 3, 7

[5] Hanting Chen, Tianyu Guo, Chang Xu, Wenshuo Li, Chunjing Xu, Chao Xu, and Yunhe Wang. Learning student networks in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6428–6437, 2021. 4

[6] Ingemar J Cox, Joe Kilian, F Thomson Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE transactions on image processing*, 6(12):1673–1687, 1997. 1

[7] Francesco Croce and Matthias Hein. Adversarial robustness against multiple and single l_p-threat models via quick fine-tuning of robust classifiers. In *International Conference on Machine Learning*, pages 4436–4454. PMLR, 2022. 5

[8] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002. 5

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 4, 6

[10] European Parliament and Council of the European Union. General data protection regulation (gdpr): Regulation (eu) 2016/679 on data protection and free movement, 2016. Available at https://gdpr-info.eu/. 1

[11] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International conference on machine learning*, pages 1607–1616. PMLR, 2018. 5

[12] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[13] Yunjie Ge, Qian Wang, Baolin Zheng, Xinlu Zhuang, Qi Li, Chao Shen, and Cong Wang. Anti-distillation backdoor attacks: Backdoors can really survive in knowledge distillation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 826–834, 2021. 8

[14] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 5

[15] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1563–1580, 2022. 2

[16] T Gu, B Dolan-Gavitt, and S BadNets. Identifying vulnerabilities in the machine learning model supply chain. In *Proceedings of the Neural Information Processing Symposium Workshop Mach. Learning Security (MLSec)*, pages 1–5, 2017. 2, 6

[17] 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, 2024. 2, 3, 6

[18] 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. 6

[19] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018. 5

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 5

[21] Mathew Hogan, Yan Michalevsky, and Saba Eskandarian. Dbreach: Stealing from databases using compression side channels. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 182–198. IEEE, 2023. 1

[22] Junyuan Hong, Yi Zeng, Shuyang Yu, Lingjuan Lyu, Ruoxi Jia, and Jiayu Zhou. Revisiting data-free knowledge distillation with poisoned teachers. In *International Conference on Machine Learning*, pages 13199–13212. PMLR, 2023. 7

[23] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1

[24] Zonghao Huang, Neil Zhenqiang Gong, and Michael K Reiter. A general framework for data-use auditing of ml models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1300–1314, 2024. 2, 6

[25] Jinyuan Jia, Yupei Liu, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of nearest neighbors against

data poisoning and backdoor attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9575–9583, 2022. 4

[26] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 4

[27] Xiao Jin, Pin-Yu Chen, Chia-Yi Hsu, Chia-Mu Yu, and Tianyi Chen. Cafe: Catastrophic data leakage in vertical federated learning. *Advances in Neural Information Processing Systems*, 34:994–1006, 2021. 1

[28] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[29] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 6

[30] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 8

[31] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*, 2021. 3, 7

[32] Yiming Li, Yang Bai, Yong Jiang, Yong Yang, Shu-Tao Xia, and Bo Li. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. *Advances in Neural Information Processing Systems*, 35:13238–13250, 2022. 2, 3, 6

[33] Yiming Li, Mingyan Zhu, Xue Yang, Yong Jiang, Tao Wei, and Shu-Tao Xia. Black-box dataset ownership verification via backdoor watermarking. *IEEE Transactions on Information Forensics and Security*, 18:2318–2332, 2023. 1, 2, 6

[34] Yudong Li, Xianxu Hou, Zheng Dezhi, Linlin Shen, and Zhe Zhao. Flip-80m: 80 million visual-linguistic pairs for facial language-image pre-training. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 58–67, 2024. 8

[35] Gaoyang Liu, Tianlong Xu, Xiaoqiang Ma, and Chen Wang. Your model trains on my data? protecting intellectual property of training data via membership fingerprint authentication. *IEEE Transactions on Information Forensics and Security*, 17:1024–1037, 2022. 2, 3, 6

[36] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pages 273–294. Springer, 2018. 3, 7

[37] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 1

[38] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. 8

[39] Ekdeep Singh Lubana, Eric J Bigelow, Robert P Dick, David Krueger, and Hidenori Tanaka. Mechanistic mode connectivity. In *International Conference on Machine Learning*, pages 22965–23004. PMLR, 2023. 6

[40] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 5

[41] Pratyush Maini, Eric Wong, and Zico Kolter. Adversarial robustness against the union of multiple perturbation models. In *International Conference on Machine Learning*, pages 6640–6650. PMLR, 2020. 5

[42] Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. Dataset inference: Ownership resolution in machine learning. In *International Conference on Learning Representations*, 2021. 2, 3, 6

[43] Xiaofeng Mao, Yuefeng Chen, Shuhui Wang, Hang Su, Yuan He, and Hui Xue. Composite adversarial attacks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8884–8892, 2021. 5

[44] NBC News. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent, 2019. Available at https://www.nbcnews.com/news/all/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921. 1

[45] OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence, 2024. Available at https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. 6

[46] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4954–4963, 2019. 5

[47] Lu Pang, Tianlong Sun, Huan Ling, and Changyou Chen. Backdoor cleansing with unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12218–12227, 2023. 7

[48] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 4

[49] 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. 4

[50] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Radioactive data: tracing through training. In *International Conference on Machine Learning*, pages 8326–8335. PMLR, 2020. 2, 6

[51] 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, 2023. 1

[52] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1

[53] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 4

[54] Ali Shafahi, Mahyar Najibi, Zheng Xu, John Dickerson, Larry S Davis, and Tom Goldstein. Universal adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5636–5643, 2020. 5

[55] Shuo Shao, Yiming Li, Mengren Zheng, Zhiyang Hu, Yukun Chen, Boheng Li, Yu He, Junfeng Guo, Tianwei Zhang, Dacheng Tao, et al. Databench: Evaluating dataset auditing in deep learning from an adversarial perspective. *arXiv preprint arXiv:2507.05622*, 2025. 3

[56] Yucheng Shi, Mengnan Du, Xuansheng Wu, Zihan Guan, Jin Sun, and Ninghao Liu. Black-box backdoor defense via zero-shot image purification. *Advances in Neural Information Processing Systems*, 36:57336–57366, 2023. 3, 7

[57] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. *Advances in neural information processing systems*, 31, 2018. 5

[58] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011. 7

[59] Salvatore J Stolfo, Malek Ben Salem, and Angelos D Keromytis. Fog computing: Mitigating insider data theft attacks in the cloud. In *2012 IEEE symposium on security and privacy workshops*, pages 125–128. IEEE, 2012. 1

[60] Bing Sun, Jun Sun, Wayne Koh, and Jie Shi. Neural network semantic backdoor detection and mitigation: A causality-based approach. In *Proceedings of the 33rd USENIX Security Symposium. USENIX Association, San Francisco, CA, USA*, 2024. 8

[61] Sebastian Szyller and N Asokan. Conflicting interactions among protection mechanisms for machine learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15179–15187, 2023. 3

[62] Sebastian Szyller, Rui Zhang, Jian Liu, and N. Asokan. On the robustness of dataset inference. *Transactions on Machine Learning Research*, 2023. 3

[63] Ruixiang Tang, Qizhang Feng, Ninghao Liu, Fan Yang, and Xia Hu. Did you train on my dataset? towards public dataset protection with cleanlabel backdoor watermarking. *ACM SIGKDD Explorations Newsletter*, 25(1):43–53, 2023. 1, 2

[64] Florian Tramer and Dan Boneh. Adversarial training and robustness for multiple perturbations. *Advances in neural information processing systems*, 32, 2019. 5

[65] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019. 2, 6

[66] Rolf Van Wegberg, Samaneh Tajalizadehkhoob, Kyle Soska, Ugur Akyazi, Carlos Hernandez Ganan, Bram Klievink, Nicolas Christin, and Michel Van Eeten. Plug and prey? measuring the commoditization of cybercrime via online anonymous markets. In *27th USENIX security symposium (USENIX security 18)*, pages 1009–1026, 2018. 1

[67] Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. Mobileclip: Fast image-text models through multi-modal reinforced training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15963–15974, 2024. 6

[68] Emily Wenger, Xiuyu Li, Ben Y Zhao, and Vitaly Shmatikov. Data isotopes for data provenance in dnns. *Proceedings on Privacy Enhancing Technologies*, 2024(1), 2024. 2, 3, 6

[69] Wikipedia contributors. Cambridge analytica — wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Cambridge_Analytica, 2024. 1

[70] Nuo Xu, Qi Liu, Tao Liu, Zihao Liu, Xiaochen Guo, and Wujie Wen. Stealing your data from compressed machine learning models. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2020. 1

[71] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. 1, 8

[72] Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *International Conference on Learning Representations*, 2022. 3, 7

[73] Yi Zeng, Minzhou Pan, Himanshu Jahagirdar, Ming Jin, Lingjuan Lyu, and Ruoxi Jia. Meta-Sift: How to sift out a clean subset in the presence of data poisoning? In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 1667–1684, Anaheim, CA, 2023. USENIX Association. 2, 3, 7

[74] Yi Zeng, Minzhou Pan, Hoang Anh Just, Lingjuan Lyu, Meikang Qiu, and Ruoxi Jia. Narcissus: A practical clean-label backdoor attack with limited information. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 771–785, 2023. 2, 6

[75] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 8

[76] Hongyu Zhu, Sichu Liang, Wentao Hu, Li Fangqi, Ju Jia, and Shi-Lin Wang. Reliable model watermarking: Defending against theft without compromising on evasion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10124–10133, 2024. 2, 5, 6

[77] Hongyu Zhu, Sichu Liang, Wentao Hu, et al. Stealing knowledge from auditable datasets. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, 2025. 3

[78] Wei Zong, Yang-Wai Chow, Willy Susilo, Joonsang Baek, Jongkil Kim, and Seyit Camtepe. Ipremover: A generative

model inversion attack against deep neural network fingerprinting and watermarking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7837–7845, 2024. 7

[79] Zihang Zou, Boqing Gong, and Liqiang Wang. Anti-neuron watermarking: protecting personal data against unauthorized neural networks. In *European Conference on Computer Vision*, pages 449–465. Springer, 2022. 2, 3, 6

[80] Chaoshun Zuo, Zhiqiang Lin, and Yinqian Zhang. Why does your data leak? uncovering the data leakage in cloud from mobile apps. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 1296–1310. IEEE, 2019. 1