Evading Data Provenance in Deep Neural Networks

Supplementary Material

Contents

Organization of the Supplementary Material

This section provides an overview of the supplementary material. The core implementation of the Escaping DOV framework can be accessed via the **Anonymous Repository**: https://github.com/dbsxfz/EscapingDOV.

The supplementary material is organized as follows:

- **Section A**: A detailed overview of the Escaping DOV framework, including illustrative examples, implementation pipelines, and pseudo-code.
- **Section B**: Descriptions and parameter configurations for DOV and baseline evasion methods.
- **Section C**: Additional experimental results and visualizations, covering: (1) weaker evasion attacks used in previous DOV literature, (2) results of SOTA evasion attacks on all DOV methods, (3) Escaping DOV on copyright datasets with significant distribution shifts from the gallery set, (4) evaluations across model architectures, (5) the side benefits of Escaping DOV, (6) time complexity analyses, and (7) examples of transfer set curation.
- **Section D**: A rigorous analysis of why Escaping DOV successfully evades all DOV methods and why other SOTA evasion attacks fail against certain DOV approaches, particularly distillation-based ones.
- **Section E**: A discussion on DOV methods in other modalities and tasks, such as large language models (LLMs), and potential countermeasures against Escaping DOV.

A. Detailed Algorithms of Escaping DOV

A.1. Zero-shot Prompt Learning

As outlined in Section 3.2 of the main text, relying solely on a text template containing class names (e.g., "a photo of class name") is insufficient for vision-language models (VLMs) to effectively capture the foreground semantics necessary to distinguish between target task categories [26]. Moreover, few-shot adaptation on the copyright dataset $\mathcal{D}$ inevitably introduces verification behaviors into the VLM [1], which impairs reliable transfer set selection.

To address these limitations, we adopt zero-shot prompt learning [26] to adapt the VLM to the target task by leveraging unbiased world knowledge embedded in a large language model (LLM). Specifically, we generate tailored image-prompts as a description set $Desc_{c_i}$ for each image category c_i using GPT-4o-mini [14], enhancing the performance of VLM.

The zero-shot prompt learning process is simplified into two steps, as illustrated in Figure 5. The distinctions between LLM-generated prompts for the LLM and image-prompts for the VLM are further detailed in Table 8.

A.1.1. Pipeline of Generating Image-prompts for VLM

1. Constructing ***LLM-prompts for LLM***: Manually craft a set of class-name-only general text templates, referred to as LLM-prompts, which are designed to elicit descrip-

Table 8. Difference Between ***LLM-prompts for LLM*** and ***Image-prompts for VLM***.

Aspect	***LLM-prompts for LLM***	***Image-prompts for VLM***
Purpose	Instruct the LLM on how to describe a category.	Serve as input to the VLM for zero-shot classification.
Content	Generalized templates with a placeholder for category names	Descriptive and category-specific sentences
Generator	Designed manually	Generated by the LLM based on the corresponding prompt templates.
Customization Level	Uniform templates used for multiple categories.	Specific to the category being described.
Examples	"Describe what a/an {category} looks like" "How can you identify a/an {category} + ?" "What does a/an {category} look like?"	"A cat has four legs, a tail, and fur." "A bird generally has wings and feathers, and can fly." "A dog is a four-legged mammal of the family Canidae."

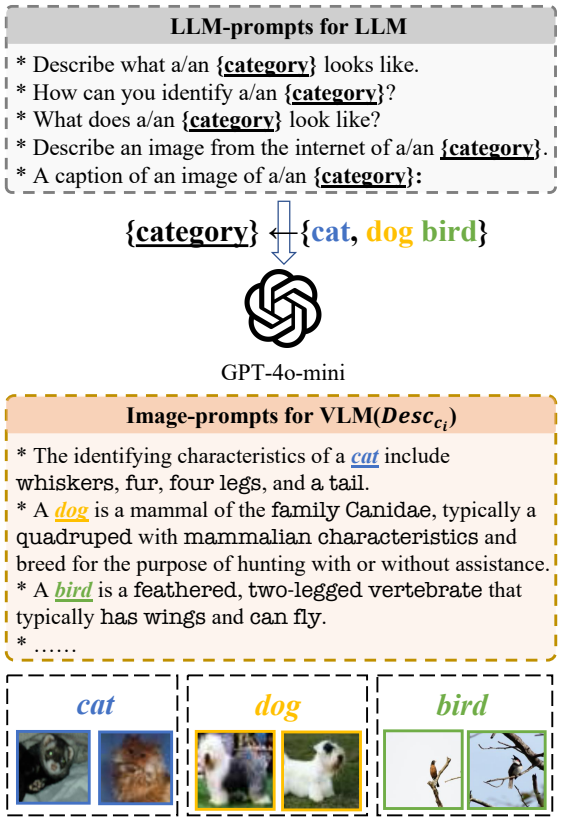


Figure 5. Pipeline of Generating Image-prompts for VLM.

tive information about a given category.

2. Generating ***Image-prompts for VLM*** using LLM : Input each LLM-prompts, filled with the specific category name c_i, into the LLM(such as GPT-4o-mini in our paper). The LLM generates multiple descriptive image-prompts $Desc_{c_i}$ for VLM that provide detailed visual descriptions of the category c_i.

A.1.2. Examples of Prompts

LLM-prompts for LLM

* "Describe what a/an {category} looks like."
* "How can you identify a/an {category}?"
* "What does a/an {category} look like?"
* "Describe an image from the internet of a/an {category}."
* "A caption of an image of a/an {category}:"

Image-prompts for VLM

- When {category} = cat, generated Image-prompts are:
  * "A cat has four legs, a tail, and fur."
  * "A typical housecat is small and has four legs."
  * "The identifying characteristics of a cat include whiskers, fur, four legs, and a tail."
  * "A cat is a small carnivorous mammal."
  * "The most common domestic cat is the brown tabby."
  * .....
- When {category} = bird, generated Image-prompts are:
  * "The bird has a long neck, short legs, and a long, thin beak."
  * "A bird is a feathered, two-legged vertebrate that typically has wings and can fly."
  * "Birds are a type of vertebrate animal, characterized by feathers, toothless beaked mouths, the laying of hard-shelled eggs, a high metabolic rate, a four-chambered heart, and a strong yet lightweight skeleton."
  * "Some identifying characteristics of a bird are that they have wings, feathers, and a beak."
  * "Most birds have wings, feathers, and beaks."
  * .....
- When {category} = dog, generated Image-prompts are:
  * "Dogs are playful, friendly, and loyal animals."
  * "A dog is a mammal of the family Canidae, typically a quadruped with mammalian characteristics and breed for the purpose of hunting with or without assistance."
  * "Some identifying characteristics of a dog are that they are a mammal, have four legs, a tail, and bark."
  * "Identifying characteristics of a dog include four legs, a tail, and fur."
  * "Most dogs have four legs, a tail, and fur."
  * .....
- When {category} = ...

A.2. Transfer Set Curation

As illustrated in Figure 1 of the main text, the transfer set curation process comprises three primary steps: (1) The VLM assigns each gallery set sample in $\mathcal{G}$ to a class t from the copyright dataset $\mathcal{D}$, thereby partitioning the gallery set $\mathcal{G}$ into K bins corresponding to the target task. (2) Within each bin, gallery samples are sorted in ascending order based on their distances to the distribution digest Cent_t.

(3) From each bin, samples are selected sequentially from the front of the sorted list. Only those samples for which the teacher model f_{θ_t} predicts the same class t are included, until the number of selected samples matches the class count $|\mathcal{D}_t|$ in the copyright dataset $\mathcal{D}$.

This process ensures two key properties of the resulting transfer set $\mathcal{T}$: (1) For all samples in $\mathcal{T}$, the teacher model f_{θ_t} and the VLM produce consistent class predictions. This consistency implies that the teacher's predictions align, at least partially, with the inherent semantics of the input images, as endorsed by the VLM. It also rules out predictions driven by specific verification behaviors (e.g., backdoor watermark outputs on trigger samples). (2) Building on the first property, samples are prioritized based on their proximity to the corresponding distribution digest. This ensures that the selected samples are more representative of the target task distribution in $\mathcal{D}$. Together, these two properties render the transfer set both *reliable* and *informative*, facilitating effective task-oriented yet identifier-invariant knowledge transfer from the teacher model f_{θ_t} to the student model f_{θ_s}.

To formally describe the transfer set curation process, the corresponding algorithm is presented in Algorithm 1. Detailed implementation details can be found in the Anonymous GitHub repository.

A.3. Selective Knowledge Transfer

To filter out suspicious verification knowledge that reflects inherent (e.g., fingerprints) or artificial (e.g., watermarks) biases in the copyright dataset $\mathcal{D}$, we propose a Selective Knowledge Transfer framework for extraction task-oriented yet identifier-invariant knowledge from the teacher model f_{θ_t} to the student model f_{θ_s}.

This process begins by generating worst-case perturbations, as described in Algorithm 2, and constructing corruption chains, detailed in Algorithm 3, which are designed to mislead the teacher model f_{θ_t} on $\mathcal{D}$. These perturbations and corruption chains are then incorporated into the distillation process, encouraging the surrogate student model f_{θ_s} to develop invariance to such misleading biases. The algorithm of the knowledge transfer process is provided in Algorithm 4.

B. Details of DOV Methods and Evasion Attacks

B.1. DOV Settings

This section describes the settings of DOV methods targeted by the proposed Escaping DOV framework. These methods are categorized into three groups: *backdoor watermarks*, *non-poisoning watermarks*, and *dataset fingerprints*.

Algorithm 1 Transfer Set Curation

Require: Copyright dataset $\mathcal{D}$, gallery set $\mathcal{G}$, teacher model f_{θ_t}, visual encoder $E_v(\cdot)$ and text encoder $E_t(\cdot)$ of the pre-trained VLM, number of $\mathcal{D}$'s classes K
Ensure: Transfer set $\mathcal{T}$

```
T ← ∅
bins ← {B_1, B_2, ..., B_K} ← ∅
for each class c_t ∈ {c_0, ..., c_K} from D do
    {Calculate the density centroid for class c_t}
    Cent_{c_t} ← (1/|D_{c_t}|) Σ_{i∈D_{c_t}} E_v(i)
    {Generate descriptions for c_t}
    Desc_{c_t} ← LLM description generation(c_t)
    r_{c_t} ← (1/|Desc_{c_t}|) Σ_{i∈Desc_{c_t}} E_t(i)
end for
{Assign images to bins based on VLM prediction}
for I in G do
    c_max ← argmax_c(sim(E_v(I), r_{c_t}))
    B_{c_max} ← B_{c_max} ∪ {I}
end for
for B_{c_t} in bins do
    {Sort images in each bin by visual similarity}
    key_func(I) ← sim(E_t(I), Cent_{c_t})
    B_{c_t} ← sorted(B_{c_t}, key = key_func, descending)
    {Filter images in each bin}
    counter ← 0
    for I in B_{c_t} do
        l_T ← f_{θ_t}(I)
        if l_T = c_t then
            T ← T ∪ {I}
            counter ← counter + 1
        end if
        if counter = |D_{c_t}| then
            break
        end if
    end for
end for
```

B.1.1. Backdoor Watermarks

1. **Badnets** [9]: A classic poison-label backdoor watermark that uses a checkerboard-style trigger (size 3×3 for CIFAR-10 and 5×5 for Tiny-ImageNet). The trigger flips the label of affected samples to a target class, causing misclassification. The poison rate (percentage of samples with triggers in $\mathcal{D}$) is set to 10%.
2. **UBW** [18]: An *untargeted* backdoor watermark, identical to Badnets in other settings, except that the labels of poisoned samples are randomized instead of being fixed to a single target class. This induces non-deterministic misclassification behavior, making it harder for backdoor defenses assuming a specific target class (e.g., Neural Cleanse [35]).
3. **Label-Consistent** [34]: A *clean-label* backdoor method that modifies only samples from the target class using

Algorithm 2 Generating Perturbations

Require: Teacher model f_{θ_t}, copyright dataset $\mathcal{D}$, number of perturbations N, number of iterations I, learning rate α, scaling factor η, regularization factor λ
Ensure: Perturbation pool $\{\delta\}$
  Initialize $\{\delta\} \leftarrow \emptyset$
  for $n = 1$ to N **do**
    $\delta_n \leftarrow 0$
    for $i = 1$ to I **do**
      for each batch $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}$ **do**
        $\hat{\mathbf{y}} \leftarrow \arg\max f_{\theta_t}(\mathbf{x})$
        $\mathbf{x}_{\text{pert}} \leftarrow \text{clip}(\mathbf{x} + \delta_n, 0, 1)$
        $\mathbf{z}_{\text{pert}} \leftarrow f_{\theta_t}(\mathbf{x}_{\text{pert}})$
        $\mathcal{L} \leftarrow \text{CrossEntropy}(\mathbf{z}_{\text{pert}}, \hat{\mathbf{y}}) - \lambda \|\delta_n\|_2^2$
        $\delta_n \leftarrow \delta_n + \alpha \cdot \nabla_{\delta_n} \mathcal{L}$
        $\delta_n \leftarrow \delta_n \cdot \min\left(1, \frac{\eta}{\|\delta_n\|_2}\right)$
      end for
      Project δ_n onto the norm (L_0, L_2, L_∞) that maximizes $\mathcal{L}$
    end for
    $\{\delta\} \leftarrow \{\delta\} \cup \{\delta_n\}$
  end for
  return $\{\delta\}$

Algorithm 3 Generating Corruption Chain

Require: Teacher model f_{θ_t}, copyright dataset $\mathcal{D}$, corruption functions $\{C_1, C_2, \ldots, C_k\}$, number of epochs N, genetic algorithms NSGA2
Ensure: Optimal corruption chain $\mathbf{s}^*$
  Initialize population p of corruption sequences $\{\mathbf{s}_i\}_{i=1}^n$
  for epoch $n = 1$ to N **do**
    Sample batch $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$
    for each sequence $\mathbf{s}_i$ in population p **do**
      {Apply corruption sequence $\mathbf{s}_i$ to $\mathbf{x}$}
      $\mathbf{x}_{\text{corr}} \leftarrow C_{s_i[1]}(C_{s_i[2]}(\ldots C_{s_i[m]}(\mathbf{x})))$
      $\mathcal{L}_i \leftarrow -\text{CrossEntropy}(f_{\theta_t}(\mathbf{x}_{\text{corr}}), \mathbf{y})$
    end for
    Update population p using NSGA2 to maximize $\mathcal{L}$
  end for
  Return the optimal corruption chain $\mathbf{s}^*$

adversarial perturbations to associate the trigger with the target class. The trigger, similar to that of Badnets, is placed in all four corners of the image. Poison rates are set to 10% for CIFAR-10 and 50% for Tiny-ImageNet to ensure verification on directly trained models.

4. **Narcissus** [41]: A *clean-label* and *invisible* backdoor watermark with a trigger constrained to a L_∞ norm of 16/255. Poison rates are 10% for CIFAR-10 and 50% for Tiny-ImageNet to ensure effective verification. The trigger is optimized on a surrogate model and acts partially as a universal adversarial perturbation, achieving slightly higher verification success rates (e.g., the VSR for the student model in Escaping DOV on Tiny-ImageNet is about 10%, comparable to unmarked models).

Algorithm 4 Selective Knowledge Transfer

Require: Teacher model f_{θ_t}, student model f_{θ_s}, perturbation pool $\{\delta\}$, corruption chain $\mathbf{s}^*$, transfer set $\mathcal{T}$, learning rate α, number of epochs N
Ensure: Optimized student model parameters θ_s
  for epoch $= 1$ to N **do**
    for each batch $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}$ **do**
      Randomly select an Operation from {Skip, Perturbation, Corruption}
      if Operation is Skip **then**
        $\mathbf{x}' \leftarrow \mathbf{x}$
      else if Operation is Perturbation **then**
        Randomly select a perturbation δ from $\{\delta\}$
        $\mathbf{x}' \leftarrow \mathbf{x} + \delta$
      else if Operation is Corruption **then**
        $\mathbf{x}' \leftarrow \mathbf{x} \circ \mathbf{s}^*$
      end if
      $y_t \leftarrow f_{\theta_t}(\mathbf{x})$ {Teacher model prediction}
      $y_s \leftarrow f_{\theta_s}(\mathbf{x}')$ {Student model prediction}
      $\mathcal{L} \leftarrow KL(y_t \parallel y_s)$ {Compute KL divergence}
      $\theta_s \leftarrow \theta_s - \alpha \nabla_{\theta_s} \mathcal{L}$ {Update student parameter}
    end for
  end for
  return θ_s

B.1.2. Non-Poisoning Watermarks

1. **Radioactive Data** [28]: Optimizes a perturbation (carrier vector) to align with the model weights trained on watermarked data. A T-test determines verification results using the lower loss on perturbed samples compared to natural samples. The perturbation is constrained to a 16/255 L_∞ norm, and the poison rate is 10%.
2. **ANW** [46]: Applies color transformations in hue space to induce lower loss on transformed samples in the watermarked model. The poison rate is 10%.
3. **Domain Watermark** [10]: Uses a domain generator to transform selected samples into a domain that is hard to generalize. This induces the model trained on watermarked data to produce higher confidence on samples from this domain. A convolutional domain generator is used, and the poison rate is 10%.
4. **Isotope** [38]: Blends external features (a fixed LSVRC-2012 image) with samples in $\mathcal{D}$ to induce high-confidence predictions on mixed verification samples. The blend ratio of original samples to external features is 0.9:0.1, and the poison rate is 10%.
5. **ML Auditor** [13] generates two sets of samples with opposing perturbations, optimizing them to be close in the input space but distant in the feature space of a surrogate model. One set is publicly released, while the other is kept private. Hypothesis testing is then performed by

comparing the target model's losses on these two sets. The poison rate is 10%.

B.1.3. Dataset Fingerprints

1. **Dataset Inference** [23]: Uses black-box adversarial perturbations to measure the distance from a sample to the decision boundary. Samples from the training set have higher distances compared to external samples, and this property is verified via a T-test. The method uses 100 samples each from the training and testing sets, requiring a total of 60,000 queries to generate adversarial perturbations.
2. **MeFA** [19]: Builds a meta-classifier for membership inference to distinguish training samples from external ones. Verification results are aggregated across 100 samples as per the original settings.

An example of watermarked samples for all methods is shown in Figure 6. It is evident that the trigger patterns are both *exclusive* and *subtle*, meaning that (1) they rarely appear in out-of-distribution (OOD) gallery sets and are therefore rarely activated by samples in the transfer set; (2) they do not alter the primary semantics of the images, while the verification behavior is orthogonal to the underlying distribution. This enables the surrogate student that avoids overfitting to specific biases to effectively mitigate such verification behavior.

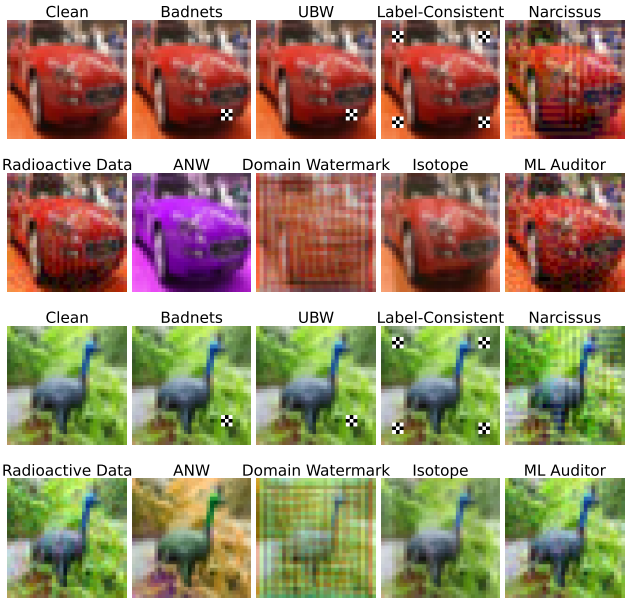


Figure 6. Visual Samples of Different Watermarking Techniques.

B.2. Evasion Attacks against DOV

1. **Fine-pruning** [20] iteratively prunes and fine-tunes the trained model to remove embedded backdoors. In our experiments, 50% of the parameters are pruned, and 5000 samples from LSVRC-2012 are used as the fine-tuning set.
2. **Meta-Sift** [40] sanitizes the training set $\mathcal{D}$ to create a smaller, clean subset for training. The clean set is set to be 50% of the size of the original training set in our experiments.
3. **Differential Privacy** [2] limits the contribution of individual samples to the model during training, thereby reducing the influence of poisoned samples. The privacy budget is set to $\epsilon = 2$.
4. **I-BAU** [39] employs universal adversarial training with implicit gradients to unlearn backdoor triggers from a suspicious model. We use 5000 samples from LSVRC-2012 as the unlearning set and adhere to the settings in the original paper.
5. **ZIP** [30] purifies test inputs using a pretrained diffusion model. We adopt guided diffusion [5] for purification and follow the original paper's configuration.
6. **NAD** [17] fine-tunes the original model as a teacher and *initializes the student with parameters from the original model*. It removes backdoors by aligning intermediate features between the teacher and student using an attention pooling loss.
7. **BCU** [25] applies an adaptive layer-wise weight dropout strategy and prediction confidence screening during distillation to suppress backdoors. The dropout rates for each layer block are set to 0.1, 0.1, 0.2, 0.3, 0.5, and 0.5, respectively, and the confidence threshold is set to 0.9.
8. **ABD** [12] integrates backdoor detection [3] with unlearning [39] to suppress backdoors during distillation. We follow the detailed settings provided in the original paper.
9. **IPRemoval** [45] combines generative model inversion with data-free distillation and employs virtual ensemble distillation to remove model watermarks. We follow the original paper's settings in our implementation.

C. Additional Experimental Results

C.1. Weaker Baseline Evasion Attacks Utilized in Previous DOV Research

In this section, we present the performance of the weaker evasion attacks used in prior DOV research to evaluate the robustness of DOV methods, as shown in Table 9. **These weaker baseline evasion attacks prove entirely ineffective against the majority of DOV techniques, underscoring why the robustness of DOV methods has often been overestimated in previous studies**. Given that these attacks fail to induce meaningful evasion in most (if not all) DOV methods and are thus not directly comparable to our Escaping DOV attack, we report their results here rather than in the main text. A detailed analysis follows below:

Methods based on *regularization* to prevent data provenance, such as the L_2 regularization in MeFA [19], label smoothing in ANW [46], distillation as well as zero-shot

Table 9. Weak Evasion Baselines Considered in Previous DOV Literature. VSR $> 30\%$ and p-value < 0.01 Indicate **Detection**, Otherwise Successful **Evasion**.

Method	Badnets		Narcissus		Isotope		Dataset Inference	
	ACC(↑)	VSR(↓)	ACC(↑)	VSR(↓)	ACC(↑)	p-value(↑)	ACC(↑)	p-value(↑)
AutoAugment	93.61	100.00	93.24	67.01	94.39	7.12e-03	94.53	1.69e-03
Gaussian Blur	90.57	100.00	89.87	55.09	90.29	1.03e-01	90.28	7.60e-05
Adv Training	84.99	100.00	85.15	10.04	84.89	2.18e-01	87.01	9.82e-05
L2 Regularization	94.33	100.00	94.52	96.33	94.80	3.00e-05	95.76	7.18e-03
Label Smoothing	93.80	100.00	94.33	93.86	94.33	7.03e-03	91.72	3.87e-03
Distillation	94.09	100.00	93.65	78.01	94.40	9.73e-03	94.29	9.98e-04
Zero-Shot Knowledge Distillation	90.46	60.74	91.56	52.13	93.59	8.33e-03	93.78	2.27e-03

knowledge distillation (ZSKD) in Dataset Inference [23], fail to evade any representative DOV techniques. While these methods occasionally reduce verification confidence, they can also amplify detection signals, as observed when L_2 regularization is applied to the Narcissus watermark. Since watermarks are carefully designed spurious features, and fingerprints are highly sensitive to confidence discrepancies between training and test data, these regularization methods are insufficient to achieve ideal generalization—i.e., preventing memorization of spurious features while maintaining nearly identical performance on training and test samples. Consequently, they fail to circumvent any DOV methods.

Notably, both distillation and ZSKD, as considered in Dataset Inference [23], employ a knowledge transfer framework similar to our Escaping DOV. However, distillation uses the original copyright dataset as the transfer set, which essentially functions as a form of label smoothing. As a result, the mapping of watermark samples to their corresponding watermark labels is preserved, and the excessive memorization of copyright training data is entirely inherited by the student model, rendering this approach totally ineffective. ZSKD, on the other hand, employs generative adversarial training to invert training data from the teacher model, maximizing the prediction difference between teacher and student to synthesize data for distillation. However, this adversarial objective inadvertently reinforces the transfer of watermark behaviors and over-memorization—e.g., many watermark trigger patterns are even directly synthesized into the transfer set—aligning with observations in [12] and experimental results in [45], as discussed in Section D.2 of the Appendix. Consequently, ZSKD achieves only marginal evasion effects in the verification process.

Methods based on *input perturbation*, which aim to *disrupt watermark patterns* during training, include data augmentation (we use AutoAugment [4], a representative method that combines multiple data augmentation operations), adversarial training as considered in ANW [46], and Gaussian blurring as explored in Isotope [38]. These methods occasionally succeed in bypassing certain DOV techniques, particularly invisible clean-label watermarks. This is because, for clean-label watermarks, the mapping between watermark triggers and watermark behaviors relies heavily on the precise presentation of the trigger pattern, while invisible watermarks generally lack sufficient resilience to input perturbations. However, these *input perturbation* approaches are *completely* ineffective against visible watermarks (e.g., BadNets) and fingerprints (e.g., DI and MeFA), rendering them unsuitable as a universal evasion baseline—i.e., an adversary with no prior knowledge of the specific DOV method employed by the copyright owner cannot reliably use them. Moreover, despite their occasional evasion success on certain DOV methods, adversarial training and Gaussian noise perturbations introduce significant performance degradation, further limiting their practical applicability.

C.2. Additional Results of SOTA Evasion Attacks on Other DOV Methods

In Section 4.3 of the main text, four representative DOV methods are selected. BadNets is a seminal work, while Narcissus, Isotope, and Dataset Inference are the most robust methods in their respective categories—backdoor watermarks, non-poisoning watermarks, and fingerprints—against our Escaping DOV attack. These methods exhibit the highest VSR or the lowest p-value after evasion in Table 1 of the main text, while the detection signal of the student models after evasion on other DOVs show no significant difference compared to independent models that have not used the copyright data.

We present the results of all SOTA evasion strategies on the remaining seven DOV methods in Table 10. Despite selecting the most advanced and powerful backdoor mitigation and privacy-enhancing techniques for comparison, our Escaping DOV remains superior in both evasion and generalization performance. Notably, it is the only method that successfully escapes all 11 DOVs. In contrast, I-BAU, ABD, and IP-Removal fail on Narcissus and/or Isotope (as shown in Table 2 of the main text), with significantly weaker generalization performance. Further in-depth analyses of

Table 10. Evasion Attacks on other DOVs. VSR $> 30\%$ and p-value < 0.01 Indicate **Detection**, Otherwise Successful **Evasion**.

Method	UBW		Label-Consistent		Radioactive Data		ANW		Domain Watermark		ML Auditor		MeFA	
	ACC(↑)	VSR(↓)	ACC(↑)	VSR(↓)	ACC(↑)	p-value(↑)	ACC(↑)	p-value(↑)	ACC(↑)	p-value(↑)	ACC(↑)	p-value(↑)	ACC(↑)	p-value(↑)
Fine-pruning	85.60	77.04	86.69	55.10	87.45	0.5303	86.83	0.8120	86.74	0.7299	85.72	1.9e-10	86.66	0.1164
Meta-Sift	86.85	12.17	87.03	82.28	87.35	0.5565	86.64	0.9665	88.48	0.9999	83.61	0.6471	88.70	0.0006
Differential Privacy	87.22	63.33	92.73	96.31	93.56	0.8248	91.15	0.0078	92.89	1.5e-23	93.06	6.2e-27	92.97	1.0e-17
I-BAU	89.07	6.46	89.31	16.46	90.37	0.1484	92.68	0.3496	84.16	0.1129	90.98	4.9e-03	89.51	0.9993
ZIP	83.92	64.70	82.57	79.80	83.13	0.0129	83.04	0.1758	83.81	0.3698	78.79	7.4e-05	83.26	0.0004
NAD	86.76	91.83	92.23	62.29	92.16	0.0942	92.16	0.0014	90.27	8.8e-06	89.40	2.3e-05	91.56	0.0005
BCU	92.26	2.21	93.17	17.24	93.46	0.4798	92.91	0.8316	92.81	**1.0000**	91.63	0.4081	93.15	0.0014
ABD	82.62	5.28	82.92	11.41	88.23	0.2167	92.58	0.1300	83.36	0.9998	85.51	0.6230	88.26	0.9996
IPRemoval	82.77	3.29	84.92	4.18	84.96	0.4234	84.14	**1.0000**	82.91	0.9999	88.18	0.7517	85.47	0.9999
Escaping DOV (Ours)	**93.41**	**1.74**	**93.19**	**3.74**	**94.07**	**0.9450**	**93.91**	**1.0000**	**93.90**	**1.0000**	**93.45**	**0.7831**	**93.93**	**1.0000**

why our Escaping DOV successfully evades all types of DOVs and why other SOTA evasion techniques fail against certain DOVs are provided in Section D.1 and D.2 of the Appendix.

C.3. Copyright and Gallery datasets with Large Distribution Shift

The primary experiments presented in the main text are conducted in scenarios where the copyright datasets are CIFAR-10 and Tiny-ImageNet, with the gallery set being LSVRC-2012 (ImageNet-1K). Alough ImageNet-1K contains images corresponding to some CIFAR-10 classes, it is not a superset of CIFAR-10. For instance, the "deer" class in CIFAR-10 lacks any direct counterpart in ImageNet-1K. **To establish a more challenging setting with no semantic overlap between the gallery set and the copyright set, we remove all 200 overlapping classes from ImageNet-1K (the gallery set) when conducting experiments on Tiny-ImageNet**. Even under this stringent condition, Escaping DOV continues to demonstrate strong performance.

Since CIFAR-10, Tiny-ImageNet, and LSVRC-2012 are all natural image datasets, **we further investigate an even more challenging scenario where the copyright dataset consists of data from specific, hard-to-obtain vertical domains, while the gallery set remains the easily accessible natural images from LSVRC-2012**. In Table 7 of the main text, we present results for two copyright datasets with distributions that are entirely distinct from the gallery set (LSVRC-2012): the face recognition dataset RAFDB and the medical diagnosis dataset OrganCMNIST. In both cases, Escaping DOV maintains strong performance. Additionally, in Table 11 of the Appendix, we extend our evaluation to four other copyright datasets that exhibit substantial distributional differences from the gallery set: the traffic sign dataset GTRSB [31], the facial emotion dataset FER2013 [8], and two datasets featuring artistic and non-photorealistic styles, ImageNet-R [11] and ImageNet-Sketch [36]. Given that ImageNet-Sketch contains only 50 samples per class, we sample 100 classes to ensure that the directly trained teacher model achieves meaningful generalization capacity.

Across all datasets, Escaping DOV evades detection successfully with minor degradation in generalization. Notably, a ResNet-18 trained on the original ImageNet-1k attains test accuracies of only 33.2% and 39.69% on ImageNet-R and ImageNet-Sketch, respectively, which is over 30% lower than the performance of the student model obtained through Escaping DOV. This highlights the advantages of training on stolen target domain data and underscores the effectiveness of our knowledge transfer approach. We attribute this effectiveness to two key factors: (1) the adaptability of knowledge distillation to intermediary transfer sets [7], and (2) the role of Transfer Set Curation (TSC) in Escaping DOV. For instance, while LSVRC-2012 does not contain a class directly corresponding to CIFAR-10's "deer" category, TSC effectively selects samples with highly similar semantic features as surrogates (see Figure 9a and Section C.7 in the Appendix).

Furthermore, these vertical-domain copyright datasets typically contain a limited number of samples per class. To ensure effective data provenance with the original (teacher) model, we apply a high watermark rate of 10% (and 50% for Narcissus). This results in the number of watermark samples often exceeding the number of clean samples in the target class. Consequently, the verification success rate (VSR) is slightly higher, and the p-value is slightly lower after our evasion attack compared with other datasets. In real-world scenarios where copyright owners are unable to watermark such a large volume of data, the evasion effectiveness of our Escaping DOV is expected to be even more pronounced.

C.4. Why Choose OOD Data as the Transfer Set in Escaping DOV

In addition to carefully selecting a transfer set from a large-scale OOD gallery using Transfer Set Curation (TSC), another possible approach for constructing the transfer set in Escaping DOV could involve using a small amount of clean IND data (if available), as demonstrated by baseline evasion attacks such as I-BAU [39] and NAD [17] do. However, we now analyze the advantages of curating a transfer set from a large-scale OOD gallery.

Firstly, due to privacy concerns and time-sensitive constraints, even a small amount of clean in-distribution (IND) data can be difficult to obtain, such as the latest medical images of new COVID-19 variants. Our Escaping DOV makes no assumptions regarding the availability of clean IND data

Table 11. Escaping DOV on Copyright Sets with Large Distribution Shift from the Gallery Set (**ImageNet-1k**). VSR $> 30\%$ and p-value < 0.01 Indicate **Detection**, Otherwise Successful **Evasion**.

Copyright Set		Badnets		Narcissus		Isotope		Dataset Inference	
		ACC(↑)	VSR(↓)	ACC(↑)	VSR(↓)	ACC(↑)	p-value(↑)	ACC(↑)	p-value(↑)
GTRSB	Vanilla	98.12	100.00	98.33	47.85	98.63	0.0040	98.68	0.0052
	Evasion	95.23	3.20	94.84	9.76	95.67	0.4935	96.34	0.1130
FER2013	Vanilla	66.62	100.00	68.01	92.15	67.71	0.0006	68.12	4.90e-08
	Evasion	62.19	7.51	62.94	12.43	63.31	0.1867	62.59	0.1429
ImageNet-R	Vanilla	65.33	100.00	66.90	57.29	66.23	0.0023	66.73	1.88e-18
	Evasion	63.47	1.92	64.60	4.23	63.97	0.4032	63.83	0.1094
ImageNet-Sketch	Vanilla	83.30	81.24	85.66	53.09	86.25	0.0069	86.05	2.88e-06
	Evasion	78.98	12.18	81.93	9.38	83.69	0.5192	82.91	0.0789

(as demonstrated in Section C.3, where we show that a fixed natural image gallery set is applicable to any distribution of the copyright dataset). **In fact, while a small amount of clean IND data is insufficient to perform an evasion attack, it can certainly enhance the effectiveness of our Escaping DOV**.

We provide two examples to support this: (1) We re-tested NAD [17] on CIFAR-10 using 5% clean IND samples. However, this did not lead to a significant improvement in performance, as NAD still failed on 6 out of 11 DOVs. (2) Taking GTRSB as an example, when 5% clean IND data was used in Escaping DOV, the student model achieved an accuracy of only 72.67%, which is over 20% lower than when the transfer set was curated using TSC from the OOD LSVRC-2012. However, when both OOD transfer set curation and the additional 5% clean IND data were used together, the student accuracy further improved to 97.18%.

C.5. Escaping DOV with Advanced Backbones

In the main text, we evaluate Escaping DOV using the relatively small ResNet-18 backbone. Here, we extend our experiments to larger and more advanced backbones, namely EfficientNet v2 [33] and Swin Transformer v2 [21]. As shown in Table 12, Escaping DOV continues to perform effectively, successfully evading all DOV methods on advanced model backbones without additional parameter tuning, despite that larger models, such as the Swin Transformer v2, exhibit severe overfitting on CIFAR-10.

C.6. Robustness as a By-product

C.6.1. Adversarial Robustness

While the test accuracy of the surrogate student is slightly lower than that of the teacher, the Selective Knowledge Transfer process inherently filters out spurious features and mitigates overfitting to undesired biases in the training data. As shown in Table 13, this process significantly enhances the surrogate student's robustness against mild adversarial perturbations, such as FGSM and PGD, particularly in the case of convolutional networks.

C.6.2. Corruption Robustness

Similar to adversarial robustness, we evaluate the teacher and surrogate student in Escaping DOV against common corruptions (e.g., Gaussian noise, shot noise (poisson noise), and impulse noise) in Table 14. Notably, the surrogate student exhibits substantial improvements over the teacher. Interestingly, the Swin Transformer, which performs worst on clean test data and under adversarial perturbations, shows the highest improvements under corruption settings, becoming the best-performing model in these scenarios. This demonstrates that the Escaping DOV framework not only evades dataset ownership verification but also significantly benefits the surrogate student in challenging settings. Consequently, this framework could be leveraged to enhance robustness against both adversarial perturbations and common corruptions, even in cases where dataset ownership verification is not a primary concern.

C.7. Illustration of Transfer Set Curation

In Figure 8a, we randomly select five images from each class in CIFAR-10, with the fifth column displaying images containing Badnets triggers. Figure 8b shows the top images from the LSVRC-2012 gallery set that are most similar to the corresponding distribution digest. We observe the following: (1) Retrieved images close to the distribution digest are visually similar to original CIFAR-10 samples, indicating that the distribution digests effectively encapsulate the task distribution and serve as reliable prototypes. (2) Notably, there is no class in LSVRC-2012 directly related to the 'deer' class in CIFAR-10. However, images from LSVRC-2012 that are closest to the 'deer' distribution digest exhibit similar features, such as horns and four legs, as seen in the 'hartebeest' and 'gazelle' classes. These images serve as effective intermediaries for knowledge transfer. (3) The unbiased VLM does not recognize the trigger pattern (a

Table 12. Escaping DOV Across Model Backbones on CIFAR-10.

DOV	ResNet-18				Efficientnet v2				Swin Transformer v2			
	Vanilla		Evasion		Vanilla		Evasion		Vanilla		Evasion	
Poisoning DOV	ACC	VSR	ACC(↑)	VSR(↓)	ACC	VSR	ACC(↑)	VSR(↓)	ACC	VSR	ACC(↑)	VSR(↓)
Badnets	94.44	100.00	93.46	1.36	94.36	100.00	92.51	1.07	90.57	100.00	89.78	1.56
Narcissus	94.76	87.34	94.37	4.59	94.85	89.93	93.58	1.29	91.36	92.71	89.98	2.47
Non-Poisoning DOV	ACC	p-value	ACC(↑)	p-value(↑)	ACC	p-value	ACC(↑)	p-value(↑)	ACC	p-value	ACC(↑)	p-value(↑)
Isotope	94.75	2.87e-03	93.99	2.84e-01	95.2	1.11e-03	93.37	1.24e-01	91.17	7.61e-03	89.96	9.51e-02
Dataset Inference	94.83	1.87e-03	93.97	4.76e-01	94.88	1.75e-05	93.35	1.47e-01	91.23	6.86e-03	90.07	1.23e-01

Table 13. Robust Accuracy of Teacher and Student Models in Escaping DOV against Adversarial Perturbations.

Model		ACC	FGSM_{L_∞} ($\epsilon = 1/255$)	PGD_{L_∞} ($\epsilon = 1/255$)	PGD_{L_2} ($\epsilon = 0.2$)
ResNet-18	Teacher	94.83	57.42	36.76	33.16
	Student	93.97	63.06 (+5.64)	51.17 (+14.41)	45.90 (+12.74)
EfficientNet v2	Teacher	94.88	58.05	42.57	37.30
	Student	93.35	68.17 (+10.12)	63.39 (+20.82)	57.29 (+19.99)
Swin Transformer v2	Teacher	91.23	45.96	33.81	31.30
	Student	90.07	47.11 (+1.15)	38.39 (+4.58)	34.23 (+2.93)

Table 14. Robust Accuracy of Teacher and Student Models in Escaping DOV against Corruptions.

Model		ACC	Gaussian Noise ($\sigma = 0.1$)	Shot Noise ($c = 50$)	Impulse Noise ($p = 0.09$)
ResNet-18	Teacher	94.83	52.78	58.48	61.60
	Student	93.97	62.31 (+9.53)	65.50 (+7.02)	63.74 (+2.14)
Efficientnet v2	Teacher	94.88	64.55	67.73	68.22
	Student	93.35	73.00 (+8.45)	74.76 (+7.03)	71.05 (+2.83)
Swin Transformer v2	Teacher	91.23	65.42	67.18	64.82
	Student	90.07	79.35 (+13.93)	80.42 (+13.24)	77.66 (+12.84)

black-and-white chessboard) as a key feature. None of the top retrieved gallery images display similar trigger patterns, thereby preventing the activation of verification behaviors during the knowledge transfer process.

However, some ambiguous samples arise when relying solely on the distribution digest criterion. For example, the top retrieved image for the 'cat' class is a 'weasel,' which visually resembles both cats and dogs. This poses a potential risk of semantic backdoor watermarks. The consensus voting mechanism between the VLM and the teacher successfully excludes such samples, as illustrated in Figure 9b, resulting in a final transfer set that is both *informative* and *reliable*.

C.8. Time Complexity of Escaping DOV

Despite involving multiple steps, our Escaping DOV framework remains computationally efficient due to the use of the *feature bank* and the *perturbation pools*. For instance, the Transfer Set Curation for CIFAR-10 takes approximately one minute, while generating the perturbation pool and corruption chains requires less than a minute. Moreover, the Selective Knowledge Transfer (SKT) module introduces negligible overhead during student model training.

To assess computational efficiency, we evaluated Escaping DOV on CIFAR-10 using a single RTX 4090 GPU, comparing it against NAD [17] and IPRemoval [45], both of which adopt similar knowledge transfer frameworks. The average runtime for Escaping DOV, NAD, and IPRemoval was 716s, 784s, and 1362s, respectively. Overall, Escaping DOV achieves approximately 10% lower runtime than ABD and 50% lower runtime than IPRemoval [12], while delivering significantly better generalization and evasion performance.

D. Framework Insight

D.1. Why Escaping DOV Successfully Evades All Types of DOVs

In general, all **watermarks** (both backdoors and non-poisoning watermarks) require the activation of the predefined *trigger* in the marked model to manifest watermark behavior. Thus, during the knowledge transfer process in our Escaping DOV framework, watermark behavior can only be transferred to the student model when the marked teacher model exhibits (at least to some extent) the watermark behavior, which can be triggered by either hard labels or probability-based soft labels). Notably, watermark triggers in DOV are inherently **exclusive**. This exclusivity arises because many copyright owners use DOV methods to protect their data, where the watermark trigger functions as a *private key*. Consequently, these triggers cannot share the same pattern; only specific trigger patterns that can authenticate the identity of the owner can be associated with their ownership. As a result, the out-of-distribution (OOD) gallery set—beyond the control of the copyright owner—is highly unlikely to contain any patterns directly linked to the watermark trigger, as shown in Appendix Figures 6 and 9a. Furthermore, the watermark trigger must be **subtle**, as the DOV watermark needs to avoid detection by both human inspection and data sanitization while ensuring it is not activated by any clean sample lacking the trigger. This is crucial to prevent performance degradation in authorized use cases (such as academic use). **Therefore, the probability of any sample in the OOD gallery (transfer) set unintentionally activating the watermark behavior is minimal**.

Moreover, consider a watermarked model whose behavior on normal samples is *identical to* that of an unrelated model (e.g., one trained on a different dataset drawn from the same IID distribution but without any copyright sam-

ples). In such cases, the student model, which mimics the watermark model's behavior on clean OOD samples without triggers, can only learn the benign behavior exhibited by the unrelated model. During the Transfer Set Curation (TSC) process, we ensure that the hard labels output by the marked teacher model are consistent with those from the CLIP model (i.e., the unmarked CLIP model endorses the teacher's hard label output). *However, the soft label outputs from the teacher model may still implicitly contain information related to the watermark behavior, and the Selective Knowledge Transfer (SKT) module within our Escaping DOV framework is specifically designed to mitigate this risk*. It generates a series of perturbations and corruptions on the teacher model that induce output changes. During the knowledge transfer process, SKT encourages the student model's invariance to these perturbations and corruptions, thereby preventing watermark-related side-channel information from being transferred to the student model via soft labels. As demonstrated in Section 4.4.2 of the main text, without the SKT mechanism, the watermark behavior in strong clean-label watermarks such as Narcissus could be transferred to the student model via soft labels. **However, SKT significantly suppresses this occurrence**.

Fingerprints fundamentally exploit the memorization (overfitting) of the copyright training data by the target model, where the unauthorized trained model exhibits significantly higher confidence (or lower loss) on the training data compared to unseen test data, providing a strong signal for data provenance. However, since the student model in Escaping DOV has never directly seen the original copyright data, its "memory" of the data is indirectly derived from the teacher model, with knowledge transfer occurring through an OOD transfer set entirely unrelated to the original copyright data. Furthermore, SKT ensures that the guidance provided to the student model is subtly distinct from the teacher's output, preventing the student from fitting to spurious features or shortcut predictions that reflect overfitting in the teacher model. As shown in Figure 2 of the main text, **our Escaping DOV significantly reduces the gap between training and test losses in the student model, making it comparable to the natural loss disparity between different subsets**. As a result, fingerprints cannot extract effective copyright signals from our student models. In contrast, Figure 7 in Appendix illustrates that baseline evasion attacks, such as NAD, lack this advantageous property and even amplify the training-test loss gap in the final deployed model.

In the following section, we briefly summarize why even the most subtle and hardest-to-bypass clean-label watermarks (including clean-label backdoors and non-poisoning watermarks) still cannot resist our Escaping DOV: (1) Regardless of the watermark type (e.g., backdoor or non-poisoning, poison-label or clean-label), all watermarks rely on *binding special predictive behavior to a trigger*. When the trigger is absent, no watermarked model exhibits any watermark behavior. (2) DOV requires that triggers be *exclusive* (distinguishable from potential watermarks in other datasets) and *subtle* (sufficiently hidden to avoid easy detection and accidental activation). Since the copyright owner cannot control the OOD gallery set, *the trigger is highly likely to be absent in the gallery set, and thus also absent in the transfer set* (see Figures 6 and 9a in the Appendix)). (3) During the **Transfer Set Curation (TSC)** process, we ensure that the hard labels in the transfer set from the teacher model are consistent with the VLMs (e.g., CLIP). Additionally, since the OOD gallery set contains no trigger pattern, *the transfer set itself does not carry any watermark clues*. For example, in the CIFAR-10 dataset protected by Narcissus, the student model trained on the hard labels of the transfer set exhibits a very low VSR (1%-2%), while the test accuracy also remains below 85%. (4) During the knowledge transfer process, clean-label watermarks *can* transfer from teacher to student through soft labels. However, the **Selective Knowledge Transfer (SKT)** module effectively mitigates this by enforcing invariance to the worst perturbations and corruptions from the teacher (as shown in Figure 4 of the main text). In summary, Escaping DOV's components work together to achieve task-oriented yet watermark-free knowledge transfer.

D.2. Why other SOTA Evasion Techniques Fail Against Certain DOVs

As discussed in Section 4.3 of the main text and Section C.2 of the supplementary material, other state-of-the-art (SOTA) evasion methods, originally designed for poison defense or privacy enhancement, exhibit varying degrees of failure when applied to certain DOV methods. In this section, we analyze the limitations of each strong evasion baseline and, in particular, focus on **why other distillation-based methods employing a similar knowledge transfer framework (e.g., NAD, BCU, ABD) are less effective than our Escaping DOV**.

Most DOV methods demonstrate robustness against fine-tuning and pruning. Although Fine-Pruning [20] represents a stronger combined attack and does pose challenges to DOV methods, it remains insufficient for complete evasion. Meta-Sift [40], which selects a clean subset of the original dataset for training, struggles against various DOV methods due to the high proportion of watermark samples (10%). Even a small number of unfiltered watermark samples can induce predefined watermark behaviors. While DP-SGD [2] mitigates the influence of individual training samples, it fails to fully neutralize the cumulative effect of multiple watermark samples sharing the same trigger. Since the trigger optimized by Narcissus differs from conventional high-frequency noise, I-BAU [39], despite reduc-

ing the verification success rate (VSR), is inadequate for complete watermark removal, aligning with the original observations in Narcissus [41]. ZIP [30], which employs a diffusion model to purify any input sample and erase watermark triggers, proves effective against high-frequency noise triggers incompatible with the original dataset features (e.g., the random noise triggers used in its original paper). However, in our experiments, the triggers employed—such as the chessboard trigger from BadNets, the optimized trigger from Narcissus, and the mixed-image trigger from Isotope—possess clear semantic meaning. As a result, ZIP's purification effect is insufficient for complete evasion.

Next, we analyze why distillation-based methods that follow a similar knowledge transfer framework, including NAD, BCU, ABD, and IPRemoval, are less effective than our Escaping DOV. **Early distillation-based backdoor defenses like NAD [17] and BCU [25] typically sought to mitigate the significant generalization degradation caused by distillation from out-of-distribution samples by sharing parameters between the original (marked teacher) model and the final deployed student model**. For instance, NAD's student model directly inherits the original model parameters (while the teacher undergoes fine-tuning), and BCU employs adaptive dropout to perturb certain parameters while still retaining over half of the original model's parameters. As a result, these methods actually function as enhanced fine-tuning techniques. Notably, **in knowledge distillation theory, parameter sharing between the teacher and student has a substantial impact on the student's ability to recover the teacher's predictive behavior** [32]. The synergistic effect of parameter sharing and teacher guidance via soft labels facilitates the transfer of latent watermark behaviors as a side effect of the inherent prediction mechanism.

To quantitatively illustrate this principle, we present a similar analysis to Figure 2 in the main text, plotting the training and test loss of intermediate models obtained by interpolating between NAD's teacher and student model parameters (see Figure 7). **Unlike the trends observed in Figure 2, two key differences emerge:** (1) There is no sharp peak in loss for the interpolated models, indicating that the student and teacher share similar predictive mechanisms [22]; (2) The gap between training and test losses does not decrease and, in many cases, even gradually increases. These characteristics make these distillation-based methods ineffective at evading both trigger-based watermarks and fingerprints that exploit training-test behavioral discrepancies. **In contrast, Escaping DOV ensures that the student model is initialized entirely independently of the marked teacher model while selecting an informative and reliable transfer set from the gallery set to preserve generalization**. This independence from the marked model results in significantly stronger evasion capabilities compared to distillation-based methods that retain parameter sharing.

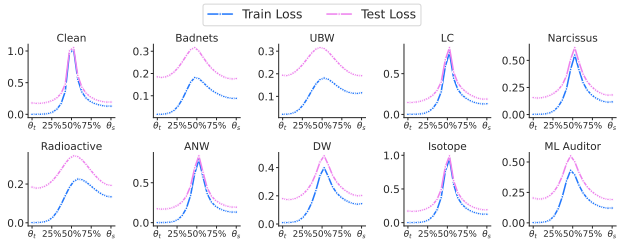
While ABD [12] does not employ parameter sharing, its distillation process is primarily tailored for backdoor removal. It filters out suspected OOD samples that could activate backdoor behaviors during distillation, thereby preventing the transfer of backdoor behavior. However, this strategy proves insufficient against advanced **non-poisoning watermarks**—such as Isotope—that do not induce misclassification behaviors. **In contrast, the selective knowledge transfer (SKT) module in our Escaping DOV framework lightly encourages the student model to be invariant to any perturbation within the input vicinity that can trigger output changes of the teacher model**. This enables it to better evade non-poisoning watermarks.

IPRemoval [45], on the other hand, utilizes generative adversarial training to invert synthetic samples from the marked model as surrogate data for distillation. However, **its synthesis objective is to maximize the predictive difference between the student and teacher models, which paradoxically increases the likelihood that watermark-related triggers remain present in the synthetic data** [12]. In contrast, our Escaping DOV employs a transfer set curation (TSC) strategy to select OOD data that is entirely beyond the copyright owner's control, ensuring a trigger-free dataset. Since IPRemoval lacks effective mechanisms to suppress the residual watermark signals embedded in its synthetic samples, it fails against the strongest watermarking schemes, such as Narcissus.

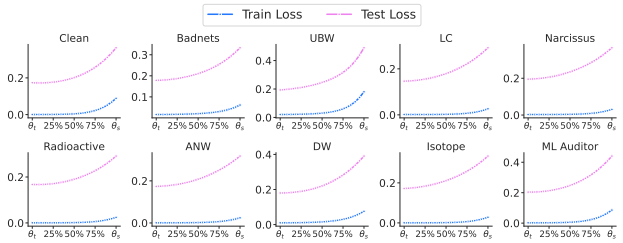
E. Extended Related Work

E.1. Concurrent Work on DOV Evasion

Shortly after the acceptance of this paper, two related studies [43] and [29] were released on arXiv, both aiming to evaluate the robustness of DOV from perspectives similar to ours. These works provide excellent analyses, though their focus differs from our *Escaping DOV*. Specifically, [43] establishes a theoretical framework that models the competition between adversary and auditor over output divergence. However, its proposed attack requires fine-tuning on pre-trained vision–language models such as CLIP, which limits its applicability to arbitrary model sizes and dataset resolutions. In contrast, [29] primarily aims to evaluate the robustness of DOV using existing techniques form other domains rather than proposing a principled new attack strategy. We believe, collectively, these three works reflect the evolving understanding of DOV robustness evaluation, and reading them together can provide broad and complementary insights for future research.



(a) Escaping DOV (Ours), Same with Figure 2 in the Main Text



(b) NAD

Figure 7. Loss Barrier Analysis of Intermediate Models between Teacher and Student Parameters in Escaping DOV and NAD: (1) Escaping DOV exhibits a sharp loss peak in interpolated models, indicating that the student model develops a distinct prediction mechanism that avoids spurious features (watermarks) inherited from the teacher [22], whereas NAD does not. (2) Escaping DOV results in a minimal training-test loss gap in student models compared to teachers, mitigating detection signals utilized by fingerprints, whereas NAD does not.

E.2. Data Provenance in Other Domains

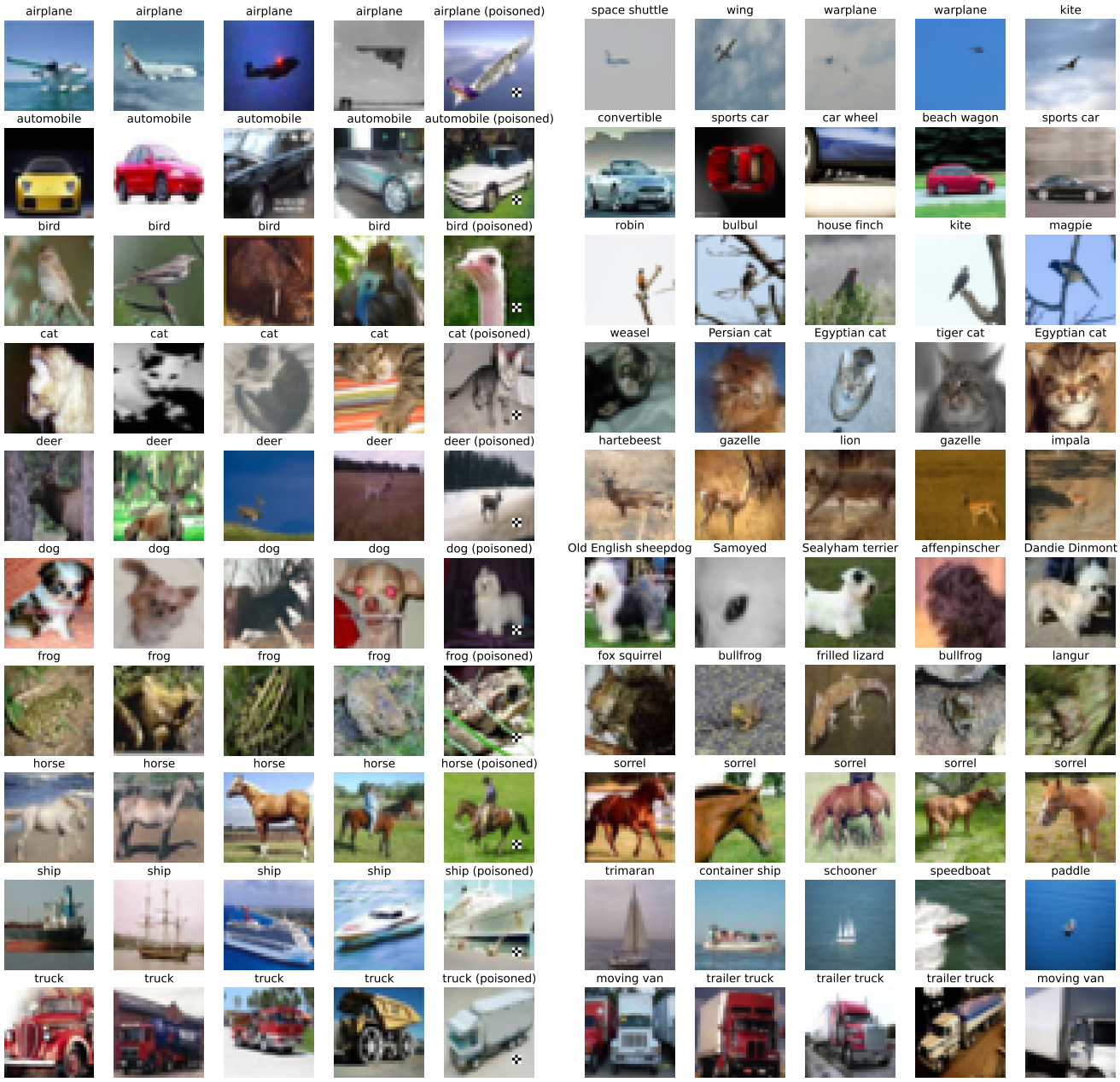
Beyond image classification, dataset ownership verification (DOV) is also a crucial concern across various modalities and tasks in deep learning. Recent studies have investigated DOV in alternative domains, including self-supervised learning [6], text-to-image diffusion models [16], 3D point clouds [37], and large language models (LLMs) [24].

DOV methods in these domains differ substantially from those designed for image classification. For instance, in the context of LLMs, the autoregressive nature lacks an explicit classification objective, and the high computational cost makes it impractical to retrain LLMs multiple times to evaluate watermark efficacy. Thus, DOV methods for LLM typically detects unauthorized data usage by analyzing the probability distribution of generated tokens to identify signals of excessive memorization. Detection signals are then aggregated across multiple text fragments with hypothesis testing to provide a robust metric for the whole dataset [24, 27, 44].

During this research, an expert suggested the SIREN watermark [16] as a potential solution for countering our evasion attack. SIREN couples watermark features with benign sample features in the representation space to detect unauthorized fine-tuning data in text-to-image diffusion models. However, its verification process depends on analyzing images generated by the diffusion model, rendering it unsuitable for classification tasks. Moreover, the fundamental differences between text-to-image generation and classification limit the direct applicability of the knowledge transfer framework in Escaping DOV for attacking the SIREN watermark in the diffusion model scenario.

Given these constraints, we explored an alternative model watermarking method, EWE [15], which similarly entangles watermark features with benign sample features within the representation space, and is applicable to classification tasks. **While EWE's watermarking process relies on controlling the teacher model's training with an additional soft nearest neighbor loss—an approach infeasible in DOV scenarios—it still fails to circumvent the evasion effect of Escaping DOV**. The resulting verification success rate (VSR) on student models consistently remains below 5%. Furthermore, in classification tasks, the observed effectiveness of feature entanglement appears to align with the *robustness pitfall* phenomenon described in [42], where the primary factor contributing to the observed effect is the increased misclassification rate from the watermark source class to the target class, rather than the successful transfer of watermark behavior.

Therefore, in future work, we aim to develop genuinely robust data provenance methods for image classification that can withstand the Escaping DOV attack. Additionally, we plan to extend Escaping DOV techniques to other domains and tasks to systematically assess the resilience of existing DOV approaches in diverse settings.



(a) CIFAR-10 (b) ImageNet

Figure 8. (a) Randomly Sampled CIFAR-10 Images, with BadNets Marked Images in the Fifth Column. (b) Top 5 Samples Closest to Class Density Centroids of Marked CIFAR-10.

References

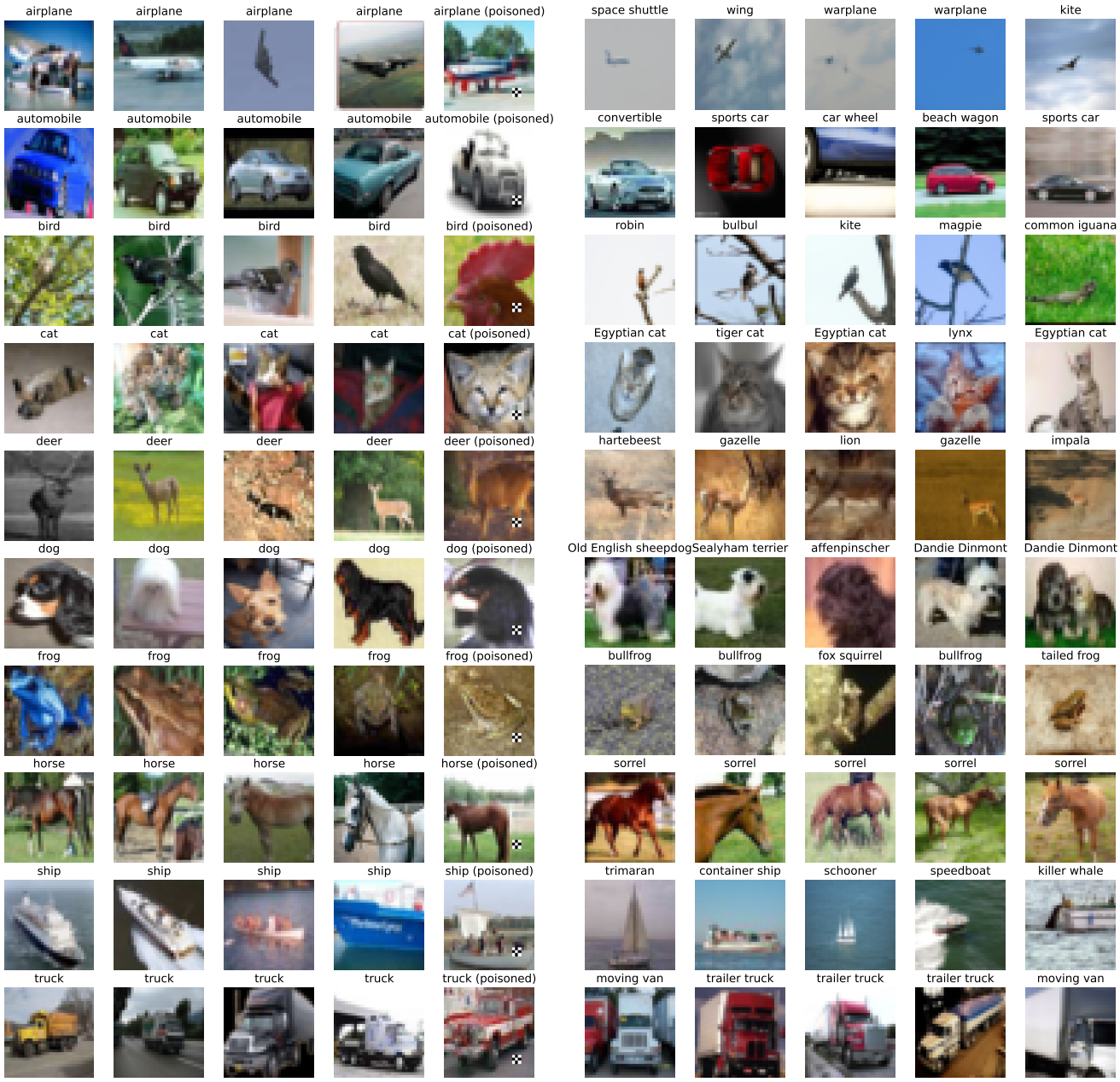
[1] Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24239–24250, 2024. 1

[2] 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. 5, 10

[3] Ruisi Cai, Zhenyu Zhang, Tianlong Chen, Xiaohan Chen, and Zhangyang Wang. Randomized channel shuffling: Minimal-overhead backdoor attack detection without clean datasets. *Advances in Neural Information Processing Systems*, 35:33876–33889, 2022. 5

[4] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019. 6

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models



(a) CIFAR-10 (b) ImageNet

Figure 9. (a) Randomly Sampled CIFAR-10 Images, with BadNets Marked Images in the Fifth Column. (b) Selected Top 5 Samples via Transfer Set Curation.

beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 5

[6] Adam Dziedzic, Haonan Duan, Muhammad Ahmad Kaleem, Nikita Dhawan, Jonas Guan, Yannis Cattan, Franziska Boenisch, and Nicolas Papernot. Dataset inference for self-supervised models. *Advances in Neural Information Processing Systems*, 35:12058–12070, 2022. 12

[7] Logan Frank and Jim Davis. What makes a good dataset for knowledge distillation? *arXiv preprint arXiv:2411.12817*, 2024. 7

[8] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Neural information processing: 20th international conference, ICONIP 2013, daegu, korea, november 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer, 2013. 7

[9] 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. 3

[10] 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. 4

[11] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 7

[12] 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. 5, 6, 9, 11

[13] 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. 4

[14] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[15] Hengrui Jia, Christopher A Choquette-Choo, Varun Chandrasekaran, and Nicolas Papernot. Entangled watermarks as a defense against model extraction. In *30th USENIX security symposium (USENIX Security 21)*, pages 1937–1954, 2021. 12

[16] Boheng Li, Yanhao Wei, Yankai Fu, Zhenting Wang, Yiming Li, Jie Zhang, Run Wang, and Tianwei Zhang. Towards Reliable Verification of Unauthorized Data Usage in Personalized Text-to-Image Diffusion Models . In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 73–73, Anaheim, CA, 2025. IEEE Computer Society. 12

[17] 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. 5, 7, 8, 9, 11

[18] 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. 3

[19] 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. 5

[20] 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. 5, 10

[21] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 8

[22] 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. 11, 12

[23] Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. Dataset inference: Ownership resolution in machine learning. In *International Conference on Learning Representations*, 2021. 5, 6

[24] Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. Llm dataset inference: Did you train on my dataset? *Advances in Neural Information Processing Systems*, 37:124069–124092, 2024. 12

[25] 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. 5, 11

[26] 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. 1

[27] Haritz Puerto, Martin Gubri, Sangdoo Yun, and Seong Joon Oh. Scaling up membership inference: When and how attacks succeed on large language models. *arXiv preprint arXiv:2411.00154*, 2024. 12

[28] 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. 4

[29] 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. 11

[30] 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. 5, 11

[31] 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

[32] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? *Advances in neural information processing systems*, 34:6906–6919, 2021. 11

[33] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 8

[34] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019. 3

[35] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE symposium on security and privacy (SP)*, pages 707–723. IEEE, 2019. 3

[36] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019. 7

[37] Cheng Wei, Yang Wang, Kuofeng Gao, Shuo Shao, Yiming Li, Zhibo Wang, and Zhan Qin. Pointncbw: Towards dataset ownership verification for point clouds via negative clean-label backdoor watermark. *IEEE Transactions on Information Forensics and Security*, 2024. 12

[38] 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. 4, 6

[39] 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. 5, 7, 10

[40] 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. 5, 10

[41] 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. 4, 11

[42] 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. 12

[43] Hongyu Zhu, Sichu Liang, Wentao Hu, et al. Stealing knowledge from auditable datasets. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, 2025. 11

[44] Hongyu Zhu, Sichu Liang, Wenwen Wang, Boheng Li, Tongxin Yuan, Fangqi Li, ShiLin Wang, and Zhuosheng Zhang. Revisiting data auditing in large vision-language models. *arXiv preprint arXiv:2504.18349*, 2025. 12

[45] 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. 5, 6, 9, 11

[46] 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. 4, 5, 6