Clean Data Can Still Carry Backdoors: Support-Persistent Backdoors for Model Reuse

Junhoo Lee[*,†] **Baekseung Kim**[*] **Seungyeon Kim** **Nojun Kwak**
Seoul National University
junhoo.lee@kaist.ac.kr
{baekseung.kim, syeonkim, nojunk}@snu.ac.kr

Abstract

Traditional backdoor attacks inject artificial trigger patterns into training data, causing models to misclassify triggered inputs while maintaining normal behavior on clean samples. Since these artificial triggers are absent in clean datasets, standard knowledge distillation (KD) typically eliminates them, leading to the common belief that KD serves as an effective purification defense. In this paper, we propose DAR (Discover-and-Relabel), a support-persistent backdoor method that retains its behavior even after clean-data distillation. DAR identifies naturally abundant patterns in low-dimensional feature subspaces using subspace clustering and applies label-only poisoning to samples that inherently satisfy the discovered rule. At test time, the backdoor is activated by locally modifying the corresponding subspace of the target images. By instantiating spatial and frequency operators, DAR achieves attack accuracy comparable to state-of-the-art backdoor attacks in ImageNet classification and CLIP-based prompt tuning. Notably, DAR keeps attack success above 96% after clean KD and remains defense-resistant in practice. Project page: https://junhoo.me/backdoor.

1 Introduction

Modern neural networks rarely remain fixed after release. Public models are often reused through knowledge distillation (KD), compression, prompt tuning, or downstream adaptation in larger systems. This lifecycle creates a practical challenge for model ownership verification. In these settings, watermarking typically relies on embedding a hidden input-output relation into a model and later querying that relation to establish provenance or detect unauthorized reuse. A key requirement is persistence: the embedded relation should remain observable even after the model is reused or adapted. A natural way to instantiate such hidden relations is through backdoor-style behavior, where a model associates an input condition with a target output. However, in a common reuse setting, a downstream user trains a student model by querying the teacher on clean, unmodified samples. This clean-query reuse makes persistence difficult because the hidden input condition may never be observed by the student.

Existing approaches face structural limitations under this clean-query reuse setting. Artificial-trigger backdoors, such as patch-,

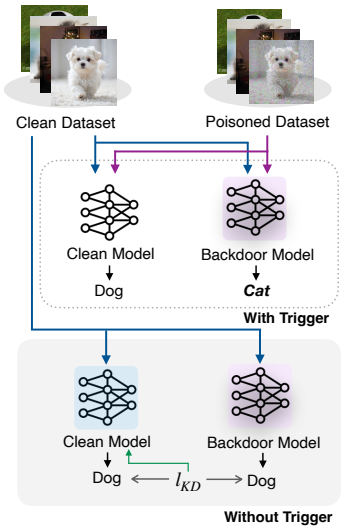


Figure 1: Failure of conventional backdoors after clean-data KD.

[*]Equal contribution.
[†]Now at KAIST.

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

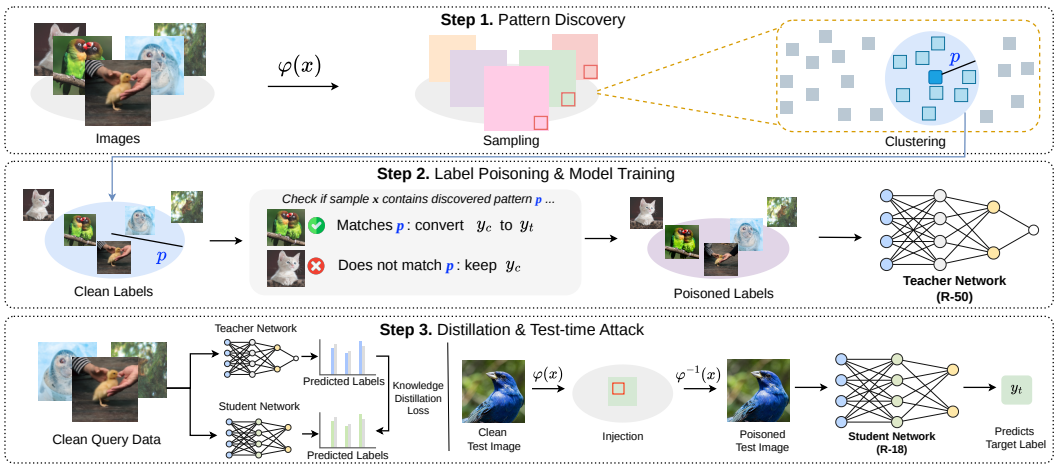


Figure 2: **DAR (Discover-and-Relabel).** DAR discovers rule-positive clean samples in a low-dimensional subspace, relabels only those samples during training, and activates the same rule at test time through the inverse subspace operation without requiring artificial trigger-positive distillation queries or external trigger exposure during downstream reuse or clean-query transfer by the student.

perturbation-, reflection-, or warping-based attacks [1–4], can achieve high attack success on the teacher, but the student cannot observe the trigger-target relation when trigger-positive inputs are absent from clean-data KD. Universal adversarial perturbations [5] provide image-agnostic directions that can be applied to arbitrary test images, but they remain optimized input artifacts that may be fragile to preprocessing or teacher-student architecture mismatch. Clean-image or label-only poisoning may appear closest to our goal. However, prior approaches primarily show that a hidden relation can be implanted into the teacher without modifying image pixels; they often rely on manually specified semantic cues, dataset-specific correlations, generative trigger optimization, or limited-scale settings [6–10]. Therefore, persistent backdoors require more than keeping poisoned training images clean. They require systematically discovering relations that are repeatedly exposed by downstream clean queries, learnable by the teacher, and controllable at test time.

Motivated by these challenges, we propose DAR (Discover-and-Relabel), a support-persistent backdoor framework for clean-data model reuse. DAR enables backdoor behavior to transfer to a downstream student using only clean queries, without exposing artificial trigger-positive samples during distillation. While conventional trigger-based attacks collapse in the student despite succeeding in the teacher, DAR maintains high attack success under the same clean-query reuse protocol. Moreover, this persistence extends to more practical model reuse settings, including OOD clean-query distillation and foundation-model adaptation.

DAR realizes this persistence through an automated Discover-and-Relabel strategy. Unlike the conventional Define-and-Enforce paradigm, which manually defines a trigger and forces training images into the trigger-positive region, DAR automatically discovers rule-positive relations that already have sufficient support in clean data and relabels only the corresponding samples. This places the malicious relation on clean data from the training stage, allowing it to be repeatedly observed through clean queries during downstream reuse. We instantiate this strategy in spatial and frequency subspaces, yielding a general rule-discovery framework that does not rely on hand-crafted semantic cues or a single trigger form. The resulting variants maintain stable persistence under clean-data distillation, input perturbations, and downstream distribution shifts.

Our contributions are as follows:

- We formulate support-persistent backdoor injection for clean-data model reuse, reframing persistence as rule injection over the clean query distribution and identifying support mismatch as the key failure mode of conventional triggers.
- We propose DAR (Discover-and-Relabel), an automated framework that mines naturally recurring patterns from clean data and relabels only the samples that already contain them, avoiding the need to expose artificial trigger samples during downstream reuse.

- We demonstrate that DAR preserves high attack success under ImageNet-scale clean-data KD, OOD distillation, cross-architecture transfer, CLIP prompt tuning, input perturbations, and standard defenses, where conventional trigger-based backdoors collapse.

2 Related Work

Trigger-Centric Backdoors. Backdoor attacks are usually studied as security threats, but the same mechanism is also used for watermarking and model ownership verification [11–13]. In both cases, the goal is to implant a controllable input-output relation while preserving normal behavior on benign inputs [1, 2, 14]. This relation is useful only if it remains observable after the model is released [15–17]. The requirement is especially important for intellectual property protection, where an unauthorized user may compress or distill a stolen model to remove identifiable behaviors while retaining task performance. Most existing attacks instantiate this control relation through a trigger-centric formulation: define a trigger pattern, modify samples to contain it, and train the model to map triggered inputs to a target label. Later variants improve stealth or robustness through blending, warping, reflections, input-adaptive triggers, color-space perturbations, or dynamic trigger generation [2, 4, 3, 18–21]. These methods differ in how the trigger is designed, but they still largely rely on externally enforced trigger-positive inputs. Consequently, high attack success on the teacher model does not inherently guarantee the persistence of the backdoor after downstream model reuse via methods such as KD.

Persistence under Model Reuse. Knowledge distillation (KD) makes this transfer-time issue explicit [22, 23]. A student trained through clean-data KD queries the teacher only on unmodified samples, so conventional trigger-positive inputs have negligible probability under the distillation data. As a result, the student observes the teacher's normal behavior and the trigger-target relation can be washed out [24–29]. This is a support-mismatch failure rather than merely a weakness of a particular trigger shape. Universal adversarial perturbations (UAPs) are relevant because they also provide image-agnostic directions that can be applied to arbitrary test images [5]. Yet UAP-style universality does not by itself establish clean-query support: the perturbation may exploit model-specific vulnerabilities and can be fragile to preprocessing or frequency filtering [30]. Persistent backdoors must be visible under clean transfer queries and activatable at inference.

Clean-Image and Label-Only Backdoors. Clean-image and label-only backdoors keep poisoned images visually unchanged by selecting or relabeling clean samples [6, 7]. They are close to our setting because they show that image modification is not necessary for teacher-side infection, but their objective is primarily training-time stealth rather than persistence under clean-query reuse. For example, FLIP [6] uses trajectory-matching optimization to synthesize poisoned labels, which requires training an expert backdoored model and weakens as benchmark scales. Natural or semantic clean-image attacks avoid synthetic triggers, but they often depend on manually specified attributes, object co-occurrences, or dataset-specific correlations [8, 31, 9]. Generative clean-image backdoors further introduce trigger optimization cost [10]. In contrast, persistent model reuse requires a rule that is automatically discovered, supported by downstream clean queries, and controllable at test time.

3 Method

We define the threat model in Section 3.1, present DAR pattern discovery in Section 3.2, and instantiate it in spatial and frequency domains for concrete clean-supported backdoors.

3.1 Problem Formulation

Threat Model. We assume the attacker has full control over the training of the Teacher Model. This includes access to the training dataset $\mathcal{D} = \{(x_i, y_i)\}$ and label manipulation capabilities, aiming to embed a backdoor targeting class y_t. In contrast, the victim trains the Student Model using only a clean dataset $\mathcal{D}_{distill}$ via Knowledge Distillation (KD). The attacker cannot intervene in the distillation process or the architecture of the Student. Because the student learns only from the support of $\mathcal{D}_{distill}$, transfer succeeds only when the trigger naturally lies in that distribution.

Challenges. Existing backdoor attacks inject triggers using predefined artificial transformations $T(x)$. However, such artificial patterns are absent in the clean dataset $\mathcal{D}_{distill}$ and are inevitably erased during distillation. Alternatively, physical trigger approaches exploit semantic objects or spurious correlations within the data distribution. Yet, these methods impose critical constraints. They typically require datasets with multi-object annotations or assume the attacker knows specific correlations within the dataset. This reliance on dataset composition [8] and prior knowledge [9] leaves automated, dataset-agnostic backdoor pattern mining unsolved in general settings and deployments.

3.2 Pattern Discovery Framework

Backdoor as Rule Injection. In clean-data KD, persistence is determined by whether the induced input-output relation has support under clean distillation queries. We denote the rule-positive region by a binary rule $\mathcal{R} : \mathcal{X} \to \{0, 1\}$, where $\mathcal{R}(x) = 1$ indicates that input x satisfies the condition that should map to the target label y_t. A backdoor can persist after reuse only if this rule is repeatedly observed during distillation and can later be activated at inference.

The Define-and-Enforce Paradigm. Based on this definition, existing attacks follow the *Define-and-Enforce* paradigm: they first define a trigger transformation T and then enforce the corresponding rule during supervised teacher training:

$$\mathcal{D}^{old}_{poison} = \{(T(x_i), y_t) \mid (x_i, y_i) \in \mathcal{D}\}. \tag{1}$$

The problem is not that the teacher cannot learn this rule; rule-positive inputs are created only by T and therefore have negligible probability under clean distillation queries. The student never repeatedly observes the trigger-conditioned behavior, so the relation has no transfer path.

The Discover-and-Relabel Paradigm. To make the relation observable under clean reuse, DAR reverses the order of design. Instead of defining a trigger and forcing inputs to contain it, we first discover an abundant natural rule $\mathcal{R}$ and relabel only the clean samples that already satisfy it:

$$\mathcal{D}^{ours}_{poison} = \{(x_i, y_t) \mid \mathcal{R}(x_i) = 1,\ (x_i, y_i) \in \mathcal{D}\}. \tag{2}$$

The poisoned samples therefore remain on the clean data support, and the target association is exposed to both teacher training and downstream clean-data KD.

Rule Discovery in a Subspace. The remaining challenge is to define a backdoor rule that is naturally supported by clean data. We choose a feature operator $\phi : \mathcal{X} \to \mathcal{F}$ and project a search set $\mathcal{S} \subset \mathcal{D}$ into the corresponding subspace:

$$\{z_1, \ldots, z_N\} = \{\phi(x_1), \ldots, \phi(x_N)\}, \quad z_i \in \mathcal{F}. \tag{3}$$

The poisoning ratio ρ specifies how many clean samples should satisfy the rule. To obtain a rule that is both common enough and specific enough, we search for the most compact region containing a ρ fraction of projected samples:

$$p^* = \operatorname*{argmin}_{p \in \{\phi(x) : x \in \mathcal{S}\}} \operatorname{quantile}_\rho\big(\{d(\phi(x), p)\}_{x \in \mathcal{S}}\big). \tag{4}$$

Given the selected prototype, the resulting rule is

$$\mathcal{R}(x) = \mathbf{1}[d(\phi(x), p^*) < t_\rho], \tag{5}$$

where $t_\rho = \operatorname{quantile}_\rho(\{d(\phi(x), p^*)\}_{x \in \mathcal{S}})$. Note that this criterion is used only to deterministically instantiate a clean-supported rule with controlled support ρ. In principle, any prototype or rule satisfying the same support and activation conditions could be used.

3.3 Rule Activation and Subspace Choices

A discovered rule is useful for a backdoor only if it can be re-entered at inference. Since our rule is defined by proximity to a prototype p^* in a subspace, activation should move a test input into the same rule-positive region while preserving the remaining semantic content. We therefore need inverse operation at test time:

$$x' = \phi^{-1}(p^*; x), \tag{6}$$

Table 1: **Clean-data KD separates teacher attack success from transferable persistence.** On ImageNet, trigger-only backdoors vanish in students, while DAR keeps high ASR across both distillation pairs.

	ResNet50 (Teacher)		ResNet18 (Student)		ResNet18 (Teacher)		ResNet18-half (Student)	
	CA↑	ASR↑	CA↑	ASR↑	CA↑	ASR↑	CA↑	ASR↑
No Attack	76.74	-	70.5	-	69.76	-	61.89	-
BadNets[1]	76.46	99.11	70.97	0.11	69.24	99	61.95	0.12
Refool[3]	75.65	97.73	70.45	0.80	68.13	96.37	61.75	0.98
WaNet[4]	76.19	99.98	70.75	0.09	69.2	98.25	61.75	0.58
Blend[2]	75.98	99.89	70.37	0.10	68.44	99.64	62.03	0.12
Spatial DAR	73.12	99.95	68.52	**96.23**	67.22	99.93	59.43	**94.83**
Frequency DAR	74.62	100	70.08	**96.65**	68.17	100	60.61	**98.96**

which replaces only the target subspace of x with the mined prototype. Note that the prototype is discovered from natural samples, and the intervention is applied only at inference to satisfy the same rule used for relabeling.

This activation requirement also determines what kind of subspace is appropriate. A useful ϕ must expose compact, high-mass regions in clean data so that the rule can be learned and transferred, while also admitting a simple inverse operation so that the rule can be imposed at test time. We instantiate this requirement in two complementary subspaces:

- **Spatial Subspace ($\phi_{spatial}$):** The most direct local test of the framework. The operator extracts a fixed region (e.g., a 4×4 patch at the bottom-right corner), where compact prototypes can be mined from clean images and later imposed by replacement.
- **Frequency Subspace (ϕ_{freq}):** Global and robust instantiation. The operator extracts the center crop of the FFT spectrum, whose low-frequency components capture image-level statistics, recur naturally across images, and can be imposed by coefficient substitution.

For the spatial domain, ϕ^{-1} pastes p^* onto the selected image region. For the frequency domain, ϕ^{-1} substitutes the central frequency coefficients with p^* and applies the inverse FFT. Thus, the rule exposed through clean relabeling and the rule activated at inference are the same rule expressed in the same subspace.

4 Experiments

This section tests support preservation in increasing scope: Section 4.1 fixes the protocol, Section 4.2 presents clean-data KD, OOD support, representation analysis, and foundation-model reuse results, Section 4.3 stress-tests robustness and defenses, and Section 4.4 validates the design choices.

4.1 Experiment Setup

Model and Datasets. To evaluate robustness on high-resolution images with diverse classes, we conduct classification experiments on ImageNet [32] rather than MNIST [33] or CIFAR10 [34]. We perform both supervised learning and knowledge distillation sequentially to measure label poisoning performance and distillability. For teacher models, we use ResNet18 and ResNet50 [35]; depending on the experiment, these teachers are distilled into ResNet18-half (ResNet18 with half channels), ResNet18 [35], EfficientNetV2-B0 [36], and DeiT-Tiny [37] students. To verify scalability, we also evaluate DAR on CLIP-based classification via prompt tuning. Some analysis and ablation experiments use ImageNet100 (100-class subset) with ResNet18 and ResNet18-half.

Baselines. For comprehensive evaluation, we classify existing attacks into four categories: (1) **Fixed-Local**, embedding small spatially concentrated patterns (patches/symbols) at fixed positions; (2) **Fixed-Global**, distributing triggers across the entire image; (3) **Fixed-Multiple**, using multiple triggers mapping to the same target; (4) **Dynamic**, generating triggers dynamically per image. Details are in the Appendix. We select four representative attacks: BadNets [1], Blend [2], Refool [3], and WaNet [4]. For defenses, we separate training/model-level mitigation from sample-level detection.

Table 2: **Support controls OOD KD transfer.** COCO/OpenGL preserve ASR; target-free SubImageNet removes it.

Distillation data	Method	CA ↑	ASR ↓
COCO	Spatial DAR	32.40	100.00
	Frequency DAR	32.30	100.00
OpenGL synthetic	Spatial DAR	15.48	94.76
	Frequency DAR	15.12	100.00
SubImageNet-50 w/ target	No Attack	88.72	-
	Spatial DAR	85.64	94.68
	Frequency DAR	87.56	99.96
SubImageNet-50 w/o target	Spatial DAR	88.60	0.00
	Frequency DAR	88.36	0.00

Table 3: **Clean-active channels explain transfer.** On ImageNet, DAR channels rank above trigger-only channels.

Method	Top-4 pct. ↑	Top-8 pct. ↑
BadNets	19.6	16.4
Blend	21.2	21.8
Refool	9.9	5.7
WaNet	0.6	1.0
Frequency DAR (T)	28.3	51.8
Frequency DAR (S)	70.6	56.5
Spatial DAR (T)	79.4	56.0
Spatial DAR (S)	45.2	42.3

The mitigation set includes fine-tuning (FT), fine-pruning (FP) [38], SAU [39], NAD [26], CLP [27], and OTBR [40]. The detection set includes STRIP [41] and Spectral [42]. Methods requiring data use 1%, following BackdoorBench [43] for consistent defense settings.

Implementation Details. For ResNet training, we use batch size 512 and learning rate 0.2 with cosine scheduling and 5-epoch warmup. Supervised learning runs for 90 epochs and knowledge distillation for 100 epochs. For DeiT-Tiny students, we use AdamW with learning rate 0.001, batch size 1024, and train for 300 epochs, following the standard DeiT recipe with strong data augmentation including Mixup, CutMix, and RandAugment. CLIP prompt tuning uses 16-shot few-shot fine-tuning per class. During distillation, teacher label temperature is 4.0 and the distillation-to-CE loss ratio α is 0.7. Backdoor attacks target a single class for all poisoned samples. For pattern identification, we use a search pool of 10,000 samples with reference set size of 10%. Poisoning rates are 5% for Spatial DAR and 3% for Frequency DAR. For evaluation, we evaluate backdoor performance using Clean Accuracy (CA) on clean data and Attack Success Rate (ASR), the target-class prediction rate on poisoned samples. For defenses, we use each method's optimal configuration for fair comparison.

4.2 Results

Persistence in Clean Data. Table 1 isolates the transfer-time failure mode. All artificial-trigger baselines reach near-perfect teacher ASR, but their student ASR drops to almost zero after clean-data KD. In contrast, DAR preserves over 96% student ASR in both teacher–student pairs. This gap shows that persistence is not explained by teacher-side attack success alone; the malicious relation must also be visible under the clean queries used to train the student. DAR is label-only poisoning: it relabels clean rule-positive samples, yet the relation transfers through KD using clean queries.

Persistence under Proxy and Partial Clean Queries. In practical model reuse, the student may distill the teacher using a public proxy set, synthetic data, or only a subset of the original domain, rather than the teacher's original training distribution. These choices can change which clean patterns are observed during KD and may remove the rule-positive samples needed for backdoor transfer. We therefore keep the poisoned teacher and discovered rule fixed, and vary only the distillation queries. As shown in Table 2, COCO and synthetic OpenGL queries reduce clean accuracy because they are weaker KD surrogates, but DAR still maintains high ASR when the discovered rule remains observable. SubImageNet-50 gives the corresponding support boundary: retaining the target class preserves transfer, while removing it eliminates the relevant clean-query support and drives ASR to zero. Thus, persistence is not guaranteed under arbitrary proxy or subset KD; it follows the support of the discovered rule under the student's query distribution.

Representation-level Analysis. To understand why DAR transfers through clean-data KD, Table 3 analyzes whether the representations involved in the backdoor are also used when the model processes ordinary clean images. We decompose the internal representation into channels and define the top-4 or top-8 channels that contribute most to the target logit on attack inputs as backdoor-related channels. We then measure how strongly these same channels are activated and contribute under 20K random

Table 5: **Label-only poisoning is sufficient.** On ImageNet/ResNet18, Frequency DAR reaches high ASR without changing training pixels.

Method	CA ↑	ASR ↓
No Attack	69.76	-
BadNets	69.24	99.00
Refool	68.13	96.37
WaNet	69.20	98.25
Blend	68.44	99.64
Spatial DAR ($\rho = 1\%$)	69.99	70.23
Spatial DAR ($\rho = 3\%$)	68.70	99.24
Spatial DAR ($\rho = 5\%$)	67.22	99.93
Freq. DAR ($\rho = 1\%$)	69.77	99.68
Freq. DAR ($\rho = 3\%$)	68.17	**100**
Freq. DAR ($\rho = 5\%$)	67.69	**100**

Table 6: **Input perturbations do not erase DAR.** Under the ResNet50 protocol, blur/JPEG collapse UAP but leave DAR ASR high.

Perturbation	UAP	DAR	
		Spatial	Freq.
None	82.90	99.95	100.00
Blur ($k = 3$)	25.94	99.98	100.00
Blur ($k = 5$)	12.91	99.95	100.00
Blur ($k = 7$)	12.76	99.94	100.00
Blur ($k = 9$)	12.75	99.94	100.00
JPEG ($q = 70$)	34.90	99.98	100.00
JPEG ($q = 50$)	9.25	99.99	100.00
JPEG ($q = 30$)	0.43	99.96	100.00
JPEG ($q = 10$)	0.18	99.35	100.00

Table 7: **Mitigation and detection comparison.** On ImageNet, defenses reduce some attacks but do not uniformly remove DAR; columns report repaired CA/ASR or clean-block/effective ASR.

Model	**Training / Model Mitigation (CA↑ / ASR↓)**					**Detection (Clean Block↓ / ASR↓)**	
	FT	FP	SAU	CLP	OTBR	STRIP	Spectral
BadNets (R50)	52.5/0.7	70.2/0.0	64.1/99.5	74.2/0.0	60.5/99.3	5.8/95.0	4.8/83.3
Blend (R50)	68.4/99.9	71.1/99.5	69.4/65.2	66.5/87.9	71.6/99.1	5.5/26.5	4.6/15.7
Refool (R50)	66.8/92.3	69.9/48.5	69.0/37.1	67.8/87.2	54.3/98.4	8.2/19.7	5.8/14.4
WaNet (R50)	68.1/0.0	71.3/0.0	70.6/0.2	77.3/13.7	69.4/90.8	5.7/79.9	4.2/92.3
Frequency DAR (R50)	69.1/99.8	68.3/100.0	68.4/100.0	73.9/99.9	60.5/100.0	6.2/38.4	3.7/100.0
Spatial DAR (R50)	71.4/35.2	65.5/43.6	64.3/99.8	78.4/0.0	64.9/100.0	6.7/83.1	3.3/99.2
Frequency DAR (R18)	54.7/3.0	63.7/97.3	61.3/65.7	69.6/97.0	70.0/96.4	3.4/95.8	4.5/93.0
Spatial DAR (R18)	56.3/21.6	63.3/50.8	60.4/95.5	70.1/0.0	68.3/96.2	6.8/82.5	3.0/94.0

clean ImageNet inputs, relative to ordinary channels. DAR backdoor-related channels have much higher clean-image contribution than conventional backdoor channels and generally lie near the range of ordinary clean channels. In contrast, conventional backdoor channels remain at very low clean-input percentiles. This indicates that conventional backdoors rely on trigger-specific representations rarely observed during clean KD, whereas DAR is embedded in representations repeatedly used by clean images, allowing the student to inherit the backdoor.

DAR to Foundation-Model Adaptation. The preceding experiments focus on ImageNet classifiers reused through clean-data KD. In practical foundation-model workflows, users typically freeze the large pretrained backbone and train only lightweight prompt or adaptation modules for the downstream task, rather than retraining or redistributing the entire backbone. This raises a different supply-chain question: can an attacker induce backdoor behavior through the adaptation module alone, without modifying the foundation model itself? To test this setting, we evaluate DAR under CLIP-based 16-shot prompt tuning. The CLIP backbone is used as the fixed foundation model, and only the downstream prompt/adaptation module is trained with a small clean ImageNet set. As shown in Table 4, Frequency DAR achieves clean accuracy and ASR comparable to the CLIP-specific backdoor baseline BadCLIP. This result suggests that DAR is not limited to KD persistence in ResNet-based classifiers; it can also induce backdoor behavior at the level of downstream adaptation modules built on top of foundation models.

Table 4: **Foundation-model adaptation.** In 16-shot ImageNet CLIP, Frequency DAR matches BadCLIP with a frozen backbone.

Method	CA	ASR
BadCLIP [44]	73.26	99.96
BadNets [1]	71.68	99.79
Frequency DAR	72.70	99.80
Spatial DAR	71.29	87.40

4.3 Robustness against Defense

Table 8: **Architecture mismatch is not the failure mode.** A ResNet50 teacher transfers DAR to both CNN and DeiT students under ImageNet KD.

Student	Method	CA↑	ASR↑
EffNetV2-B0	Base	76.69	-
	Spatial DAR	74.61	89.89
	Frequency DAR	76.73	84.64
DeiT-Tiny	Base	74.18	-
	Spatial DAR	72.48	88.42
	Frequency DAR	74.16	**96.19**

Table 9: **Small pools recover representative rules.** On ImageNet-100, support stays near the target rate.

Search Pool	Actual Poisoning Rate	
	Spatial	Freq.
0.009%	2.33	2.31
0.043%	3.78	1.97
0.086%	**5.24**	1.83
0.429%	**5.27**	**3.08**
0.857%	4.86	**3.32**
1.714%	5.12	3.74

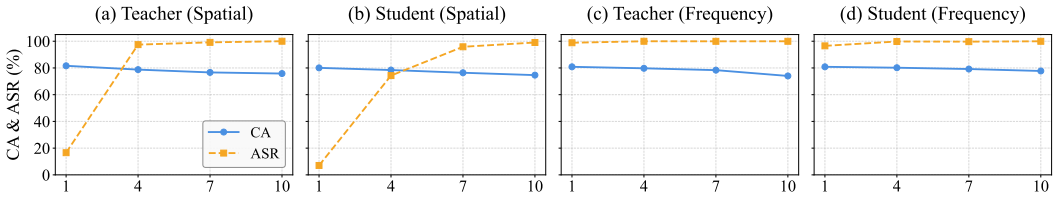


Figure 3: **Poisoning-rate controls support.** Poisoning rate vs. performance on ImageNet100 with ResNet18→ResNet18-half. Frequency DAR reaches high ASR at lower rates.

Against Input Perturbation Defenses. A natural concern is that DAR may simply behave like a universal adversarial perturbation (UAP): both apply a shared, image-agnostic modification that can induce target misclassification across many inputs. The distinction we test here is whether the attack relies on a fragile input-level perturbation or on a clean-supported rule learned by the model. To separate these cases, we apply input transformations such as Gaussian blur and JPEG compression, which weaken fine-grained perturbation patterns before inference [45–47]. If DAR were merely exploiting a subtle perturbation artifact, its ASR should degrade sharply under these transformations. However, under the same ResNet50 target-class protocol, Table 6 shows that UAP [5] rapidly collapses after blur and JPEG compression, whereas both DAR variants retain near-perfect ASR. Thus, DAR is clean-supported rather than perturbation-driven under common input sanitization.

Against Backdoor Defense. Finally, Table 7 shows that DAR remains robust under standard backdoor mitigation and detection defenses. Conventional attacks poison training images by inserting triggers or applying transformations, whereas DAR keeps the training images clean and changes only the labels of rule-positive samples. Despite this label-only poisoning constraint, Frequency DAR preserves high ASR after many defenses, often matching or exceeding conventional trigger-injection backdoors. This demonstrates the persistence of clean backdoors.

4.4 Ablation Studies

Overview. The preceding experiments show that DAR preserves high ASR after clean-data KD, and that this persistence is linked to a clean-supported rule. Here, we test whether this conclusion is an accidental outcome of a particular setting. We vary architecture, budget, trigger strength, search-pool size, target class, and KD weight across settings, ablations, and model families.

Cross-Architecture Transfer. To test whether DAR depends on a specific teacher–student architecture pair, we distill a ResNet50 teacher into students from different model families. EfficientNetV2-B0 tests transfer to a CNN variant, while DeiT-Tiny tests whether persistence is limited to ResNet-to-ResNet or CNN-family feature alignment. Table 8 shows that DAR keeps high ASR on EfficientNetV2-B0, and Frequency DAR also keeps high ASR on DeiT-Tiny. This suggests that DAR persistence is not confined to same-family distillation and can transfer to both CNN variants and ViT-style students under augmentation-heavy KD with Mixup, CutMix, and RandAugment.

Poisoning Rate. We next vary the poisoning rate to analyze performance-stealthiness tradeoff. The poisoning rate controls exposure of the rule-positive relation during training and distillation, directly tying it to the support-based explanation of DAR. As shown in Figure 3, ASR can decrease at small

Table 10: **Target choice is not special.** On ImageNet-100, clean-data KD preserves DAR ASR across multiple target classes.

Method	Target	Teacher		Student	
		CA↑	ASR↑	CA↑	ASR↑
Spatial DAR	0	78.24	88.64	78.10	81.14
	40	78.20	94.44	78.74	79.14
	80	78.68	98.58	78.10	85.82
Frequency DAR	0	79.92	99.78	80.30	97.12
	40	80.18	99.88	80.00	97.06
	80	80.14	99.84	80.16	97.56

Table 11: **KD weight controls imitation strength.** Frequency DAR keeps high ASR on ImageNet even when α is reduced.

Method	α	CA↑	ASR↑
No Backdoor	-	61.89	-
Spatial DAR	0.5	60.90	52.51
	0.6	60.33	74.14
	0.7	59.43	94.83
Frequency DAR	0.5	61.76	94.82
	0.6	61.35	93.33
	0.7	60.61	**98.96**

poisoning rates, but both DAR variants retain high ASR once enough rule-positive support is available. Thus, DAR remains effective with adequate support.

Rule Discovery with Limited Search Pools. Because DAR uses subspace clustering to find rule-positive patterns in clean data, we test whether it requires searching a large fraction of the training set. Table 9 reduces the search pool size and then measures whether the discovered rule has full-dataset support close to the target poisoning rate. Except for extremely small pools, small-subset rules remain representative on the full dataset, so DAR works with limited search budgets.

Trigger Strength. We also vary trigger strength by changing the Spatial patch area and the Frequency amplitude. This tests whether DAR depends on an overly strong or conspicuous test-time modification. Figure 4 shows that extremely small patches or weak frequency amplitudes can reduce ASR, but both variants remain effective under weaker triggers than the main setting. Thus, DAR offers a practical ASR–stealthiness trade-off while the Discover-and-Relabel principle does not hinge on one tested trigger strength.

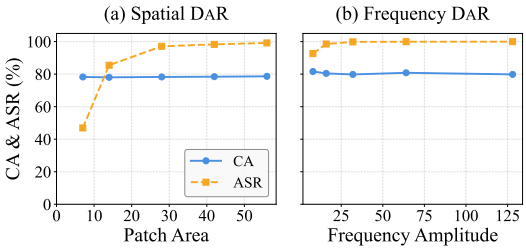


Figure 4: **Trigger strength is not a knife edge.** Trigger strength vs. backdoor performance on ImageNet100 using ResNet18. DAR remains effective.

Target-Class Robustness. We then vary the target label to check whether DAR succeeds only for an unusually easy target class. Table 10 evaluates multiple target classes on ImageNet-100. Both Spatial DAR and Frequency DAR maintain consistent teacher and student ASR, so persistence is not explained by one favorable boundary or class-specific bias across these targets.

KD Loss Weight Sensitivity. Finally, we investigate the impact of distillation parameters on transferability. Table 11 varies the KD loss weight α in the ResNet18-teacher to ResNet18-half-student setting. As α decreases, ASR drops in some settings, showing that weaker imitation of the teacher's soft behavior can weaken backdoor transfer. However, DAR does not collapse under lower KD weight: Frequency DAR remains highly persistent even at $\alpha = 0.5$, and Spatial DAR recovers as the KD signal strengthens. Thus DAR is clean-supported across KD strengths.

5 Conclusion

We introduced DAR (Discover-and-Relabel), which reframes distillation-resilient backdoors as support-preserving rule discovery. Rather than injecting artificial triggers that disappear under clean-data KD, DAR discovers naturally abundant subspace rules and relabels only clean samples that already satisfy them, making the malicious relation observable during teacher training and downstream clean-query reuse. Our experiments show that DAR consistently preserves high attack success after clean-data KD, whereas conventional artificial-trigger backdoors that succeed in the teacher largely collapse in the student. This validates the central premise of DAR: persistence depends not merely on teacher-side attack success, but on whether the malicious input-label relation

is supported by clean data and remains observable during distillation. Thus, clean-data KD should not be viewed as a general purification defense, because it can remove unsupported trigger artifacts without eliminating malicious relations embedded in clean-supported reused models.

Limitations This work studies support-persistent backdoors as a mechanism-level vulnerability in clean-data model reuse. Our empirical scope is limited to vision models, with concrete instantiations in spatial and frequency subspaces and evaluations under digital activation, clean-data KD, prompt tuning, input perturbations, and standard defense pipelines. Thus, our results should be interpreted as evidence that clean-supported malicious relations can persist in these reuse settings, rather than as an exhaustive characterization of all possible modalities, activation channels, or defender strategies.

Acknowledgments and Disclosure of Funding

This work was supported by the Korean Government through grants from IITP (RS-2021-II211343 and RS-2025-25442338) and KISA (RS-2026-25527246).

References

[1] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. doi: 10.1109/ACCESS.2019.2909068.

[2] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[3] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision – ECCV 2020, Part X*, volume 12355 of *Lecture Notes in Computer Science*, pages 182–199. Springer, Cham, 2020. doi: 10.1007/978-3-030-58607-2_11.

[4] Tuan Anh Nguyen and Anh Tuan Tran. WaNet – imperceptible warping-based backdoor attack. In *International Conference on Learning Representations (ICLR)*, 2021.

[5] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[6] Rishi D. Jha, Jonathan Hayase, and Sewoong Oh. Label poisoning is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL `https://openreview.net/forum?id=prftZp6mDH`.

[7] Dazhong Rong, Guoyao Yu, Shuheng Shen, Xinyi Fu, Peng Qian, Jianhai Chen, Qinming He, Xing Fu, and Weiqiang Wang. *Clean-Image Backdoor Attacks*, pages 187–202. Springer Nature Switzerland, 2024. ISBN 9783031723599. doi: 10.1007/978-3-031-72359-9_14. URL `http://dx.doi.org/10.1007/978-3-031-72359-9_14`.

[8] Emily Wenger, Roma Bhattacharjee, Arjun Nitin Bhagoji, Josephine Passananti, Emilio Andere, Haitao Zheng, and Ben Y Zhao. Natural backdoor datasets. *arXiv preprint arXiv:2206.10673*, 2022.

[9] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[10] Binyan Xu, Fan Yang, Di Tang, Xilin Dai, and Kehuan Zhang. Breaking the stealth-potency trade-off in clean-image backdoors with generative trigger optimization, 2025. URL `https://arxiv.org/abs/2511.07210`.

[11] 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.

[12] 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.

[13] Junhoo Lee, Mijin Koo, and Nojun Kwak. CSF: Black-box fingerprinting via compositional semantics for text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL `https://arxiv.org/abs/2604.16363`.

[14] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network and Distributed System Security Symposium (NDSS)*, pages 1–15. The Internet Society, 2018.

[15] Junhoo Lee, Hyunho Lee, Kyomin Hwang, and Nojun Kwak. Deep support vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL `https://arxiv.org/abs/2403.17329`.

[16] Dongkwan Lee, Junhoo Lee, and Nojun Kwak. Deep edge filter: Return of the human-crafted layer in deep learning. In *Advances in Neural Information Processing Systems*, 2025. URL `https://arxiv.org/abs/2510.13865`.

[17] Kyomin Hwang, Suyoung Kim, Junhoo Lee, and Nojun Kwak. Do not think about pink elephant! In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. URL `https://arxiv.org/abs/2404.15154`.

[18] Tuan Anh Nguyen and Anh Tuan Tran. Input-aware dynamic backdoor attack. In *Advances in Neural Information Processing Systems (NeurIPS) 33*, 2020.

[19] Wenbo Jiang, Hongwei Li, Guowen Xu, and Tianwei Zhang. Color backdoor: A robust poisoning attack in color space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/papers/Jiang_Color_Backdoor_A_Robust_Poisoning_Attack_in_Color_Space_CVPR_2023_paper.pdf.

[20] Jie Zhang, Dongdong Chen, Qidong Huang, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. Poison ink: Robust and invisible backdoor attack. *arXiv preprint arXiv:2108.02488*, 2021. URL https://arxiv.org/abs/2108.02488.

[21] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. In *Proceedings of the 2022 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2022. URL https://arxiv.org/abs/2003.03675.

[22] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. arXiv:1503.02531.

[23] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. doi: 10.1007/s11263-021-01453-z.

[24] Huy Phan, Jinqi Xiao, Yang Sui, Tianfang Zhang, Zijie Tang, Cong Shi, Yan Wang, Yingying Chen, and Bo Yuan. Clean and compact: Efficient data-free backdoor defense with model compactness. In *European Conference on Computer Vision*, pages 273–290. Springer, 2024.

[25] Lu Pang, Tao Sun, Haibin Ling, and Chao Chen. Backdoor cleansing with unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12218–12227, 2023. doi: 10.1109/CVPR52729.2023.01176.

[26] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *9th International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2101.05930. arXiv:2101.05930.

[27] Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. Data-free backdoor removal based on channel lipschitzness. In *European Conference on Computer Vision (ECCV)*, pages 175–191. Springer, 2022.

[28] 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.

[29] Zonghao Ying and Bin Wu. Nba: defensive distillation for backdoor removal via neural behavior alignment. *Cybersecurity*, 6(1):20, 2023.

[30] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

[31] Emily Wenger, Josephine Passananti, Yuanshun Yao, Haitao Zheng, and Ben Y. Zhao. Backdoor attacks on facial recognition in the physical world. *CoRR*, abs/2006.14580, 2020. URL https://arxiv.org/abs/2006.14580.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[33] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[34] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[35] 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 (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[36] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10096–10106. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/tan21a.html.

[37] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

[38] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Research in Attacks, Intrusions and Defenses (RAID 2018)*, volume 11050 of *Lecture Notes in Computer Science*, pages 273–294. Springer, Cham, 2018.

[39] Shaokui Wei, Mingda Zhang, Hongyuan Zha, and Baoyuan Wu. Shared adversarial unlearning: Backdoor mitigation by unlearning shared adversarial examples. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[40] Weilin Lin, Li Liu, Jianze Li, and Hui Xiong. Fusing pruned and backdoored models: Optimal transport-based data-free backdoor mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26299–26307, 2025.

[41] Yansong Gao, Chang Xu, Derui Wang, Shiping Chen, Damith C. Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference (ACSAC)*, pages 113–125. ACM, 2019. doi: 10.1145/3359789.3359790. URL https://dl.acm.org/doi/10.1145/3359789.3359790.

[42] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018.

[43] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, Mingli Zhu, Ruotong Wang, Li Liu, and Chao Shen. Backdoorbench: A comprehensive benchmark and analysis of backdoor learning. *International Journal of Computer Vision*, pages 1–88, 2025.

[44] Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Badclip: Trigger-aware prompt learning for backdoor attacks on clip, 2024. URL https://arxiv.org/abs/2311.16194.

[45] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E. Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression. *CoRR*, abs/1705.02900, 2017. URL http://arxiv.org/abs/1705.02900.

[46] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. *CoRR*, abs/1711.00117, 2017. URL http://arxiv.org/abs/1711.00117.

[47] Li-Yun Wang. Adversarial perturbation suppression using adaptive gaussian smoothing and color reduction. In *2021 IEEE International Symposium on Multimedia (ISM)*, pages 158–165, 2021. doi: 10.1109/ISM52913.2021.00033.

A Societal Impacts

This research introduces DAR (Discover-and-Relabel), which exposes a critical security vulnerability by demonstrating that backdoor triggers can survive clean-data knowledge distillation, a process previously widely considered an effective purification defense. By shifting the attack paradigm from artificial trigger injection to the hijacking of naturally abundant, task-relevant features, we show that even models trained or distilled with verified clean data are susceptible to inheriting malicious behaviors. While this persistence offers a powerful new tool for robust model watermarking and intellectual property protection, it more importantly highlights the urgent need for advanced defenses against feature-based threats in clean-data reuse pipelines.

B Unlearning Dilemma

A natural question is why Frequency DAR appears more persistent than the spatial variant, both under clean-data distillation and under several defenses. We interpret this difference through the class-wise structure of the discovered natural triggers. Figure 5 shows that spatial rules are spread more evenly across classes, whereas Frequency DAR produces a much more class-concentrated rule-positive distribution. This means that the frequency rule is not just an arbitrary perturbation direction or spurious noise pattern. Instead, it aligns with image statistics that are unevenly tied to particular semantic classes, making the rule closer to a task-relevant feature than an isolated artifact.

This interpretation explains the unlearning dilemma. During clean-data distillation, the student is trained to imitate the teacher on ordinary images. If the frequency rule is entangled with class-relevant clean features, clean queries repeatedly expose the same statistical relation that supports the backdoor. The student therefore has no clear signal that this relation should be discarded as malicious; it can be inherited as part of the teacher's legitimate decision rule. Conversely, a defense that tries to remove the frequency-supported behavior must suppress features that also help classify clean images from the affected classes. The defense results in Table 7 are consistent with this tension: Frequency DAR often retains high ASR after standard mitigation methods, while reductions in ASR are not uniformly achieved without clean-accuracy degradation or failure under other defenses. Thus, the dilemma is not simply that the backdoor is hard to erase, but that the evidence in Figure 5 suggests why erasing it can conflict with preserving useful class-correlated structure.

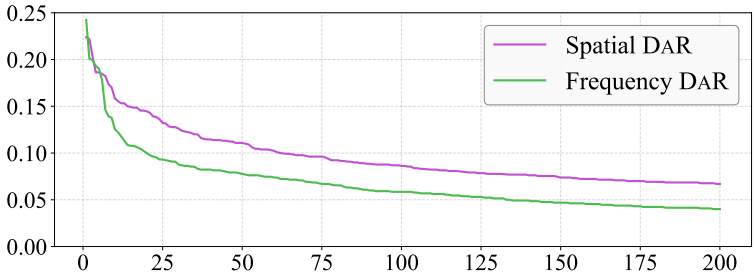


Figure 5: **Rule support is class-structured.** On ImageNet, Frequency DAR concentrates rule-positive samples more strongly than Spatial DAR.

C Additional OOD Support Measurements

The main paper summarizes OOD support in Table 2; here, ImageNet-C measurements show that the discovered rules remain numerically present under common corruptions and shifted image statistics.

Table 12: **Rule support survives common corruptions.** On ImageNet-C, weighted trigger ratios remain close to the source poisoning rates.

Corruption	Spatial	Freq.	Corruption	Spatial	Freq.
brightness	2.16	0.48	jpeg compression	6.69	2.48
contrast	11.49	25.82	motion blur	10.95	3.88
defocus blur	11.77	3.68	pixelate	7.98	2.86
elastic transform	7.30	2.74	saturate	3.86	1.97
fog	13.90	32.37	shot noise	2.60	2.98
frost	0.46	1.66	snow	1.46	0.23
gaussian blur	11.71	3.85	spatter	6.31	1.57
gaussian noise	2.47	3.51	speckle noise	4.89	2.85
glass blur	10.51	3.80	zoom blur	9.03	4.16
impulse noise	1.69	5.18		-	
Weighted Average (Overall)					
Spatial (5%): **6.70**			Frequency (3%): **5.58**		

D Additional Design Checks

We include two compact checks orthogonal to the main distillation result: replacing the default L2/MSE rule distance with L1 preserves high attack success, and grayscale evaluation keeps both DAR variants effective, indicating that the discovered relations are not restricted to color-only cues.

Table 13: **The rule is not tied to one distance or color cue.** On ImageNet/ResNet18, DAR remains effective with L1 distance and grayscale evaluation.

(a) L1 rule distance

Method	CA	ASR
No Attack	69.76	-
Spatial DAR	67.28	99.97
Frequency DAR	69.02	95.08

(b) Grayscale evaluation

Method	CA	ASR
No Attack	68.27	-
Spatial DAR	65.93	99.77
Frequency DAR	66.28	99.96

E Types of Attacks

In this section, we provide a comprehensive overview of backdoor attack taxonomies and justify our experimental setup that uses four representative attack methods: BadNets [1], Blend [2], Refool [3], and WaNet [4]. These methods cover the major backdoor attack types used in our evaluation.

F Qualitative Results

To provide a concrete understanding of our method's operation, Table 14 presents qualitative examples of the poisoning process. Specifically, the left column displays images selected as targeted poison samples according to our criteria, while the right column shows the results after injecting triggers into benign images. These examples illustrate the selection strategy for targeted poisoning and the visual characteristics of poisoned samples generated by our attack method in the appendix.

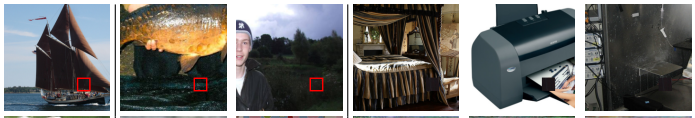
F.1 Attack Taxonomy

Backdoor attacks can be categorized by trigger type and its relationship to the input data. Based on these characteristics, we classify attacks into static and dynamic attacks for the evaluation taxonomy.

F.2 Static Attacks

Static attacks use consistent triggers and can be subdivided by spatial characteristics:

Table 14: **Selected clean positives define the test-time rule.** ImageNet examples pair mined Spatial/Frequency positives with their triggered variants.

	Poisoned samples					
Description	Ex.1	Ex.2	Ex.3	Ex.4	Ex.5	Ex.6
Spatial DAR						
Frequency DAR						

Local Trigger Attacks. **Representative method: BadNets [1]** Local trigger attacks embed a fixed patch or symbol at a chosen position. Key characteristics include consistent spatial placement:

- Fixed position, typically in a corner or less salient region of the image
- Consistent pattern across all poisoned samples
- Relatively easy to implement but also easier to detect
- Direct modification of pixel values in a localized region

BadNets exemplifies this approach by inserting a small fixed pattern, such as a square, at a consistent location. It is highly effective but leaves a distinctive signature identifiable by stronger defenses.

Global Trigger Attacks. **Representative method: Blend [2]** Global trigger attacks distribute the trigger pattern across all or most of the input image. Key characteristics include:

- Pattern spans the entire input space rather than a localized region
- Often implemented through blending or transparency effects
- More subtle visual appearance compared to local triggers
- May be less disruptive to semantic content

Blend superimposes a controlled-transparency pattern over the entire image, making it less visually obvious than patch attacks while preserving high attack success in practice across poisoned inputs.

Multiple Trigger Attacks. **Representative method: Refool [3]** Multiple trigger attacks employ several different triggers that all map to the same target label. Their key characteristics include:

- Uses multiple distinct trigger patterns
- All triggers direct classification to the same target
- Increases attack robustness against defenses that may only identify some triggers
- Often leverages naturally occurring visual elements

Refool uses reflection patterns as triggers, embedding natural-looking reflection effects that strongly influence predictions and remain robust against many defenses and detectors.

F.3 Dynamic Attacks

Representative method: WaNet [4] Unlike static attacks, dynamic attacks use input-dependent triggers that adapt to each image and its content. Their key characteristics include:

- Trigger transformation depends on the input image content
- No fixed pattern that repeats across poisoned samples
- Often implemented through subtle warping or transformations

- Significantly harder to detect through statistical analysis

WaNet implements this approach through imperceptible image warping, where the warping function creates an image-specific trigger. This makes the attack difficult to defend against because traditional methods that search for consistent poisoned-input patterns fail to identify the dynamic trigger.

G Defense Evaluation Details

We separate the defenses in Table 7 into two groups because they answer different operational questions. Training/model-level mitigation methods modify the model and are evaluated by the resulting clean accuracy (CA) and attack success rate (ASR). Detection methods leave the model fixed and instead reject suspicious inputs; for these methods, we report the clean block rate and the effective ASR after rejected attack inputs are removed.

Training and model-level mitigation. FT fine-tunes the backdoored model on a small clean subset. FP prunes low-activation channels and then fine-tunes the remaining model. SAU and NAD use clean data to unlearn the backdoor behavior through adversarial unlearning or attention distillation. CLP removes channels using channel Lipschitzness scores without relying on poisoned samples. OTBR fuses the defended model with a pruned counterpart through optimal transport. Unless a method is data-free by design, it is given 1% of ImageNet training data. For CLP, we sweep the pruning threshold and report the best valid point under a clean-accuracy constraint, avoiding degenerate settings that erase ASR only by collapsing clean accuracy.

Detection defenses. STRIP detects inputs with low prediction entropy under input superposition. Spectral scores samples through representation-space anomalies. We calibrate detection thresholds on clean validation samples at a target false positive rate of 5%. The table reports clean block rate and effective ASR after detection, so an effective detector should keep the clean block rate low while sharply reducing ASR. This reporting is distinct from CA/ASR after mitigation because detectors reject inputs rather than repair classifiers during inference-time filtering and deployment.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes].

   Justification: The abstract and introduction state the proposed DAR method, its distillation persistence claim, and the evaluation settings used to support those claims.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes].

   Justification: See limitations sections

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A].

Justification: The paper does not present formal theoretical results or proofs.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes].

   Justification: The experimental setup describes datasets, architectures, training schedules, distillation settings, trigger settings, and evaluation metrics needed to understand the reported results.

   Guidelines:

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

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

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

Answer: [Yes].

Justification: The submitted supplementary material includes the code and instructions needed to reproduce the main experiments.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes].

Justification: The experiments section reports the primary datasets, model architectures, learning rates, batch size, schedules, epochs, distillation temperature, and loss weighting.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [No].

Justification: The current tables and figures report point estimates without error bars or statistical significance tests. ImageNet-scale experiments are computationally expensive, so each main setting was run once.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes].

   Justification: Experiments were run on NVIDIA RTX A6000 GPU workers with 48GB memory. The ImageNet experiments use the same computational scale as standard ImageNet training for the corresponding architecture and schedule, with the training epochs and distillation epochs reported in the experimental setup.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes].

   Justification: The work studies backdoor vulnerabilities and defenses in a research context and discusses the security implications of the findings.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes].

Justification: The impact statement discusses both potential use for model watermarking and the security risk posed by distillation-resilient backdoors.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [N/A].

Justification: The current manuscript does not release new high-risk models or datasets.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes].

Justification: Existing assets are credited through citations and used according to their respective licenses or terms. ImageNet is used under its non-commercial research terms; COCO annotations are under CC BY 4.0 with images governed by their source terms; OpenAI CLIP is MIT-licensed; DeiT is Apache-2.0-licensed; and BackdoorBench is CC BY-NC 4.0-licensed.

Guidelines:

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

13. **New assets**

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

    Answer: [N/A].

    Justification: The current manuscript does not introduce or release new datasets, models, or other assets.

    Guidelines:

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

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

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

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or human-subject experiments.

    Guidelines:

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

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

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

    Answer: [N/A].

Justification: The paper does not involve crowdsourcing or human-subject experiments.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A].

Justification: The core method development does not use LLMs as an important, original, or non-standard research component.

Guidelines:

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