REVISITING ROBUSTNESS GAINS OF VISION-LANGUAGE MODELS UNDER LABEL-GUIDED ATTACK

Anonymous authors
Paper under double-blind review

ABSTRACT

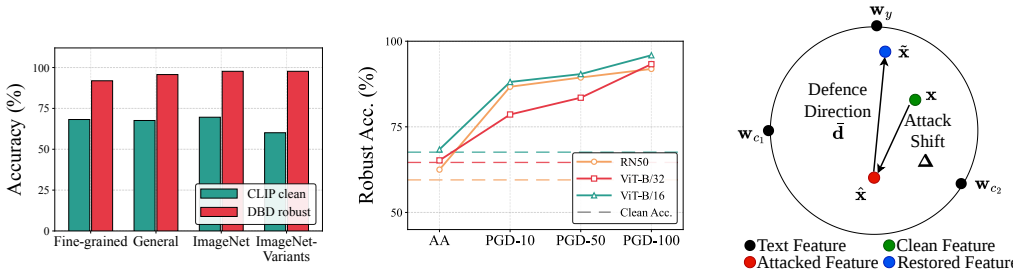
It is generally expected that, even after applying a defense, models remain less accurate on adversarial inputs than on clean inputs. However, we observe that Directional Bias-guided Defense (DBD), a test-time defense for zero-shot CLIP, achieves high adversarial accuracy under ground-truth-guided (GT-guided) attacks, even surpassing clean zero-shot CLIP. We study this counterintuitive result and identify a critical evaluation caveat: when the attack objective explicitly uses the ground-truth label, the attack process leaves recoverable label information in the attacked feature that can be exploited by a subsequent recovery mechanism. Theoretically, we analyze this result from three aspects: (1) GT-guided attacks push the features away from the true class; (2) DBD can recover such attack direction from the transformed views; and (3) after reconstructing the attacked feature along this direction, the true class margin can exceed the margin on the clean feature, accounting for why DBD can surpass clean zero-shot CLIP. Empirically, our results across multiple datasets, attack settings, and CLIP backbones support this analysis. We vary the attack-label assignment through ground-truth, pseudo, loop, fixed, and controlled-corruption settings. DBD accuracy tracks the ground-truth label proportion and changes across these assignments. We further evaluate conditional subsets, perturbation budgets, other test-time defense pipelines, and complete-pipeline adaptive attacks. Finally, we introduce **D-Pred**, a simpler direction-only prediction method that serves as a diagnostic probe of attack-induced label information. It surpasses DBD in average accuracy under GT-guided PGD-100 across all three backbones, yet approaches guessing on clean inputs, exposing attack-dependent predictive information in the recovered direction.

1 INTRODUCTION

Adversarial robustness usually concerns how to mitigate the performance degradation caused by adversarial perturbations (Szegedy et al., 2014; Wang et al., 2024a): when an adversarial perturbation disrupts a model's prediction, a defense aims to improve performance on attacked inputs and reduce the gap to clean inputs. This perspective also applies to zero-shot CLIP (Radford et al., 2021), where images are classified by matching visual features with text prototypes in a shared embedding space. Although CLIP achieves superior zero-shot recognition through large-scale image-text pretraining, it remains vulnerable to adversarial attacks (Wang et al., 2024a). Existing defenses for vision-language models, including robust optimization (Mao et al., 2023; Wang et al., 2024b; Madry et al., 2018; Zhang et al., 2026a), prompt tuning (Zhou et al., 2024; Shu et al., 2022; Yoon et al., 2024), and test-time defense (Liu et al., 2025; Zhang et al., 2026b), can improve adversarial accuracy. However, defended accuracy is generally expected to remain below the clean zero-shot baseline.

This expectation is violated by Directional Bias-guided Defense (DBD) (Liu et al., 2026), a recent test-time defense for zero-shot CLIP. Under ground-truth-guided attacks (GT-guided attacks), DBD achieves adversarial accuracy that exceeds clean zero-shot CLIP, as shown in Figure 1a. The result is not isolated: Figure 1b shows that the same pattern appears across common GT-guided attacks (Goodfellow et al., 2015; Madry et al., 2018; Carlini & Wagner, 2017; Croce & Hein, 2020). Such a result is difficult to interpret as a standard robustness gain, because the defended adversarial inputs become easier to classify than the clean inputs. This raises a key question: Does the gain come from DBD alone, or does it also depend on how the adversarial examples are generated?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Robustness of DBD (b) GT-Guided Attack Optimization (c) Direction Recovery Mechanism

Figure 1: Analysis of DBD under label-guided attacks. (a) Under PGD-100, DBD robust accuracy can exceed CLIP clean accuracy. (b) Under GT-guided PGD, DBD accuracy increases from PGD-10 to PGD-100. (c) DBD recovers an approximate reverse attack direction and extrapolates along it to improve the true-class margin.

We argue that the attack generation process is essential for interpreting this gain. In the standard GT-guided setting (Goodfellow et al., 2015; Madry et al., 2018; Carlini & Wagner, 2017; Croce & Hein, 2020), the attack objective explicitly uses the ground-truth label. As a result, the attack may not only make the image adversarial, but also leave recoverable label information in the attacked feature. DBD provides a concrete setting for analyzing this caveat: given an attacked image, it applies input transformations to estimate a direction from the transformed views, and then reconstructs the attacked feature by moving along this direction. This structure can convert attack-induced label information into prediction gain: transformed views can expose a direction related to the reverse attack shift, and reconstruction along this direction can make the true-class margin exceed the corresponding margin on the clean feature.

Theoretically, we analyze the attack–recovery process in Figure 1c from three aspects: (1) GT-guided attacks induce a label-conditioned feature shift away from the true class; (2) transformed views can align with the reverse attack shift, with low-entropy views providing a cleaner estimate; and (3) reconstruction along this direction can make the true-class margin exceed that of the clean feature. This analysis explains how DBD can surpass clean zero-shot CLIP. To test this explanation, we examine the true label's position in the attacked logit ordering, whether entropy-selected views provide a cleaner recovery direction, and how margin change relates to true-label recovery. We further vary attack-label assignments, the ground-truth label proportion, and perturbation budgets, evaluate subsets defined by clean predictions, and apply the attacks to other test-time defense pipelines.

These observations motivate **D-Pred**, a simpler direction-only prediction method that serves as a diagnostic probe of attack-induced label information. By classifying from the recovered direction alone, it surpasses DBD in average accuracy under GT-guided PGD-100 across all three backbones, yet approaches guessing on clean inputs. This contrast supports our central claim that the recovered direction carries label information that can account for DBD's gain.

Our contributions are summarized as follows:

- We analyze DBD from three aspects: GT-guided attacks push the attacked feature away from the true class, transformed views recover the reverse attack direction, and reconstruction along this direction can make the true-class margin exceed the corresponding margin on the clean feature.
- We design empirical experiments around this analysis, varying the attack-label assignment, ground-truth label proportion, conditional clean-prediction subset, and perturbation budget. We also evaluate other test-time defense pipelines and complete-pipeline adaptive attacks.
- We introduce D-Pred, a simpler direction-only predictor that probes attack-induced label information. Its higher average accuracy than DBD under GT-guided PGD-100 across all three backbones, together with near-zero clean accuracy, supports the predictive role of the recovered direction.

2 Related Work

Adversarial robustness for CLIP. CLIP (Radford et al., 2021) enables strong zero-shot transfer, yet its adversarial robustness remains limited (Szegedy et al., 2014; Wang et al., 2024a). Existing

defenses for vision-language models improve empirical robustness through adversarial training (*e.g.*, TeCoA (Mao et al., 2023), PMG-AFT (Wang et al., 2024b), FARE (Schlarmann et al., 2024)), prompt-based test-time adaptation (*e.g.*, R-TPT (Sheng et al., 2025) and TAPT (Wang et al., 2025)), and transformation-based input correction (*e.g.*, TTC (Xing et al., 2025), DOC (Jiang et al., 2025) and AOM (Tong et al., 2025)). These methods can improve accuracy under attack, but they may incur substantial inference overhead or degrade clean-image performance. DBD (Liu et al., 2026) is particularly unusual in this landscape: it identifies traces of ground-truth label information left by the attack, reconstructs new image features from those traces, and can achieve accuracy on adversarial images that substantially exceeds the clean-image baseline. Our goal is not to introduce another robustness method for CLIP, but to understand why DBD can appear unusually effective under label-guided attack and how this affects the interpretation of robustness in this setting.

Label-dependent attacks and evaluation effects. Standard adversarial attacks such as FGSM (Goodfellow et al., 2015), PGD (Madry et al., 2018), CW (Carlini & Wagner, 2017), and AutoAttack (Croce & Hein, 2020) are commonly instantiated with the ground-truth label when constructing adversarial examples. This label dependence is often treated as a routine part of the evaluation protocol, but it can also introduce hidden biases (Athalye et al., 2018) into robustness measurements. In particular, when the attack objective explicitly depends on the correct class, the resulting adversarial feature may retain structured traces (Zhou & Patel, 2024; Kurakin et al., 2017) of that label rather than being purely label-agnostic noise. Our analysis is therefore related to broader work showing that robustness evaluations can be shaped by implicit dependencies between the attack, the defense, and the evaluation pipeline (Tramèr et al., 2020; Zhao et al., 2023). The setting we study makes this interaction especially clear: the attack uses the ground-truth label, while DBD estimates a consensus recovery direction from transformed views at test time. Our contribution is to connect these two facts and show that DBD's core mechanism can exploit label information injected by the attack itself.

3 Problem Setup and Analysis

This section formalizes the setting and analyzes DBD under GT-guided attacks. We first define zero-shot CLIP and the attack setting in Section 3.1, summarize DBD in Section 3.2, analyze the recovery process in Section 3.3.

3.1 Preliminaries

Zero-shot CLIP. CLIP (Radford et al., 2021) maps images and text prompts into a shared embedding space through an image encoder f_θ and a text encoder g_ϕ. For a prompt set $\{t_c\}_{c=1}^{C}$, the normalized text embedding of class c is $\mathbf{w}_c = g_\phi(t_c)/\|g_\phi(t_c)\|$. For an input image x, the image feature is $\mathbf{x} = f_\theta(x)/\|f_\theta(x)\|$, and the logit is $s_c(\mathbf{x}) = \mathbf{w}_c^\top \mathbf{x}$. Zero-shot CLIP predicts:

$$\hat{y}_{\text{zs}}(\mathbf{x}) = \arg\max_{c\in[C]} \; s_c(\mathbf{x}). \tag{1}$$

GT-guided attacks. Following the evaluation setting of Directional Bias-guided Defense (DBD) (Liu et al., 2026), adversarial examples are generated against vanilla CLIP and then passed to the defense. We focus on cross-entropy PGD (Madry et al., 2018) as a representative GT-guided attack. Given a clean image x with label y, PGD generates an adversarial perturbation:

$$\delta^* = \arg\max_{\|\delta\|_\infty \le \varepsilon} \mathcal{L}_{\text{CE}}(x+\delta, y), \tag{2}$$

via iterative sign-gradient ascent projected onto the ℓ_∞-ball of radius ε. The attacked image is $\hat{x} = x + \delta^*$, and its feature is $\hat{\mathbf{x}} = f_\theta(\hat{x})/\|f_\theta(\hat{x})\|$. DBD is then applied to the attacked input.

Label assignment. For attack label $\ell \in [C]$, PGD generates:

$$\delta^*(\ell) = \arg\max_{\|\delta\|_\infty \le \varepsilon} \mathcal{L}_{\text{CE}}(x+\delta, \ell). \tag{3}$$

For feature $\mathbf{z}$, the gradient is $\nabla_{\mathbf{z}} \mathcal{L}_{\text{CE}}(\mathbf{z}, \ell) = \frac{\boldsymbol{\mu}(\mathbf{z}) - \mathbf{w}_\ell}{\tau}$. GT-guided attacks use $\ell = y$; pseudo-, loop-, and fixed-label attacks use alternative assignments. For n examples, the ground-truth label proportion is $p_{\text{GT}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\ell_i = y_i]$.

3.2 Overview of DBD

Directional Bias-guided Defense (DBD) (Liu et al., 2026) is a test-time defense with two phases.

Phase 1: directional consensus estimation. Starting from the attacked image $\hat{x}$, DBD applies a diverse set of transformations and extracts the corresponding features. It then uses the k views with the lowest CLIP prediction entropy, denoted $\{\hat{\mathbf{x}}_i\}_{i=1}^{k}$, and computes the mean feature shift:

$$\bar{\mathbf{d}} = \frac{1}{k}\sum_{i=1}^{k}(\hat{\mathbf{x}}_i - \hat{\mathbf{x}}), \qquad \rho = \frac{1}{k}\sum_{i=1}^{k}\left\langle \frac{\hat{\mathbf{x}}_i - \hat{\mathbf{x}}}{\|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}\|}, \frac{\bar{\mathbf{d}}}{\|\bar{\mathbf{d}}\|} \right\rangle, \tag{4}$$

where the DB-score ρ measures the directional concentration of the transformed feature shifts.

Phase 2: feature reconstruction. DBD shifts the feature along the estimated direction:

$$\tilde{\mathbf{x}} = \begin{cases} \hat{\mathbf{x}} + \bar{\mathbf{d}}, & \rho \le \eta \text{ (likely clean)}, \\ \hat{\mathbf{x}} + \lambda\bar{\mathbf{d}}, & \rho > \eta \text{ (likely adversarial)}, \end{cases} \tag{5}$$

with threshold η and amplification factor $\lambda > 1$. Equivalently, $\tilde{\mathbf{x}} = \hat{\mathbf{x}} + \beta(\rho)\bar{\mathbf{d}}$, where $\beta(\rho) = 1$ if $\rho \le \eta$ and $\beta(\rho) = \lambda$ if $\rho > \eta$. In the analysis below, $\tilde{\mathbf{x}}$ denotes the reconstructed representation in Eq. equation 5 itself, so the class score remains linear in the analyzed variable. The final prediction is $\hat{y}_{\text{DBD}} = \arg\max_c \; s_c(\tilde{\mathbf{x}})$.

3.3 Theoretical Analyses of DBD under GT-guided Attack

We use cross-entropy PGD as the main case and discuss related label-guided objectives such as CW (Carlini & Wagner, 2017) in Appendix A.1. The analysis proceeds from the attack-induced feature shift to the reverse direction estimated from transformed views, and finally to the margin gain produced by DBD's feature reconstruction.

Analysis 1: PGD attacks encode the label information in feature space. In PGD-based attacks on CLIP, the image encoder and text embeddings are fixed, and the cross-entropy objective is back-propagated through the frozen CLIP logits. For a generic CLIP image feature $\mathbf{z}$, this gives:

$$\nabla_{\mathbf{z}}\mathcal{L}_{\text{CE}}(\mathbf{z}, y) = \frac{\boldsymbol{\mu}(\mathbf{z}) - \mathbf{w}_y}{\tau}, \qquad \boldsymbol{\mu}(\mathbf{z}) = \sum_{c=1}^{C}\pi_c(\mathbf{z})\mathbf{w}_c. \tag{6}$$

The coefficient $\pi_c(\mathbf{z}) = \text{softmax}_c(\{s_j(\mathbf{z})/\tau\}_{j=1}^{C})$ is the softmax probability of class c. Eq. equation 6 shows that the CE update is determined by the gap between the softmax-weighted text embedding and the true-label text embedding. Since PGD maximizes $\mathcal{L}_{\text{CE}}$, we examine whether the ascent direction decreases the true-label logit and the pairwise margin between y and each competing class. The derivation of Eq. equation 6 is deferred to Appendix A.2.

Proposition 3.1 (Directional structure of the feature-space gradient)**.** *For normalized and distinct text embeddings, let* $\mathbf{q}_c = \mathbf{w}_y - \mathbf{w}_c$ *and define:*

$$E_c(\mathbf{z}) = \sum_{j \notin \{y,c\}} \pi_j(\mathbf{z})\mathbf{q}_c^{\top}(\mathbf{w}_j - \mathbf{w}_y). \tag{7}$$

For any feature $\mathbf{z}$*, the CE feature-space gradient satisfies:*

$$\mathbf{w}_y^{\top}\nabla_{\mathbf{z}}\mathcal{L}_{\text{CE}}(\mathbf{z}, y) < 0. \tag{8}$$

Moreover, for each $c \neq y$*, we have:*

$$\mathbf{q}_c^{\top}\nabla_{\mathbf{z}}\mathcal{L}_{\text{CE}}(\mathbf{z}, y) < 0 \quad \Longleftrightarrow \quad \pi_c(\mathbf{z})\|\mathbf{q}_c\|^2 > E_c(\mathbf{z}). \tag{9}$$

The proof is in Appendix A.2. For the final adversarial feature, we define the accumulated feature-space PGD shift as:

$$\boldsymbol{\Delta} = \hat{\mathbf{x}} - \mathbf{x}, \tag{10}$$

When the accumulated shift follows the same margin-decreasing direction, it satisfies $(\mathbf{w}_y - \mathbf{w}_c)^{\top}\boldsymbol{\Delta} < 0$ for the corresponding competing class.

Remark. Eq. equation 6 shows that the feature-space update of cross-entropy PGD depends on the true-label prototype $\mathbf{w}_y$. Consequently, the accumulated attack shift $\boldsymbol{\Delta}$ can vary with the ground-truth label, rather than behaving as a label-agnostic feature shift.

Analysis 2: transformed views recover the reverse attack direction. DBD's transformations are approximately label-preserving, so each transformed feature $\hat{\mathbf{x}}_i$ is compared with the clean feature $\mathbf{x}$. We define the average residual as:

$$\bar{\boldsymbol{\xi}} = \frac{1}{k}\sum_{i=1}^{k}(\hat{\mathbf{x}}_i - \mathbf{x}). \tag{11}$$

With this residual, the averaged DBD direction is:

$$\bar{\mathbf{d}} = \frac{1}{k}\sum_{i=1}^{k}(\hat{\mathbf{x}}_i - \hat{\mathbf{x}}) = -\boldsymbol{\Delta} + \bar{\boldsymbol{\xi}}. \tag{12}$$

If each selected view's feature is closer to the clean feature than the attacked feature is, the triangle inequality gives $\|\bar{\boldsymbol{\xi}}\| \le k^{-1}\sum_{i=1}^{k}\|\hat{\mathbf{x}}_i - \mathbf{x}\| < \|\boldsymbol{\Delta}\|$, so $r < 1$ in the following proposition.

Proposition 3.2 (Recovery with the reverse attack shift)**.** *With $\bar{\mathbf{d}}$ defined in Eq. equation 12, let $r = \|\bar{\boldsymbol{\xi}}\|/\|\boldsymbol{\Delta}\|$. If $r < 1$, then its cosine alignment with the reverse attack shift satisfies:*

$$\left\langle \frac{\bar{\mathbf{d}}}{\|\bar{\mathbf{d}}\|}, -\frac{\boldsymbol{\Delta}}{\|\boldsymbol{\Delta}\|} \right\rangle \ge \sqrt{1 - r^2} > 0. \tag{13}$$

The proof is in Appendix A.3; Section 5.3 reports r and the alignment in Eq. equation 13.

Remark. Eq. equation 13 gives a directional role to entropy filtering. By selecting low-entropy views, DBD reduces the average residual $\bar{\boldsymbol{\xi}}$ in Eq. equation 12, so the averaged direction $\bar{\mathbf{d}}$ is less affected by residual variation and more dominated by the reverse attack shift.

Analysis 3: reconstruction can exceed the clean margin. We use the true-class margin:

$$m_y(\mathbf{z}) = \min_{c \ne y}(\mathbf{w}_y - \mathbf{w}_c)^\top \mathbf{z}, \tag{14}$$

with correct classification equivalent to $m_y(\mathbf{z}) > 0$.

For $c \ne y$, let $\mathbf{q}_c = \mathbf{w}_y - \mathbf{w}_c$ and define:

$$\gamma_c = \mathbf{q}_c^\top \mathbf{x} - m_y(\mathbf{x}) \ge 0, \qquad A_c = (\mathbf{w}_c - \mathbf{w}_y)^\top \boldsymbol{\Delta}, \qquad M = \min_{c \ne y} A_c. \tag{15}$$

Using Eqs. equation 12 and equation 5, the realized reconstruction update is:

$$\tilde{\mathbf{x}} - \mathbf{x} = \boldsymbol{\Delta} + \beta(\rho)\bar{\mathbf{d}} = -\big(\beta(\rho) - 1\big)\boldsymbol{\Delta} + \beta(\rho)\bar{\boldsymbol{\xi}}. \tag{16}$$

Therefore, the margin change is:

$$m_y(\tilde{\mathbf{x}}) - m_y(\mathbf{x}) = \min_{c \ne y}\left[\gamma_c + \big(\beta(\rho) - 1\big)A_c + \beta(\rho)\mathbf{q}_c^\top \bar{\boldsymbol{\xi}}\right]. \tag{17}$$

Proposition 3.3 (Margin beyond the clean feature)**.** *Define:*

$$B = min_{c \ne y}\left[\gamma_c + \big(\beta(\rho) - 1\big)A_c - \beta(\rho)\|\mathbf{q}_c\|\,\|\bar{\boldsymbol{\xi}}\|\right]. \tag{18}$$

If $B > 0$, then $m_y(\tilde{\mathbf{x}}) > m_y(\mathbf{x})$.

The proof is deferred to Appendix A.4, and the quantities in Eqs. equation 15 and equation 18 are measured in Section 5.4. A larger positive margin places the feature farther from the nearest competing class boundary. Thus, when the reconstructed adversarial feature exceeds the clean-feature margin, it is easier to classify than the original clean feature.

4 **D**-PRED: A DIRECTION-ONLY METHOD

This motivates **D**-Pred, which uses the recovered direction as the classification feature:

$$\mathbf{z}_{\bar{d}} = \frac{\bar{\mathbf{d}}}{\|\bar{\mathbf{d}}\|}, \qquad \hat{y}_{\bar{d}} = \arg\max_{c \in [C]} \mathbf{w}_c^\top \mathbf{z}_{\bar{d}}. \tag{19}$$

Because $\hat{y}_{\bar{d}}$ depends only on $\bar{\mathbf{d}}$, **D**-Pred serves as a diagnostic probe of attack-induced label information in the recovered direction. High accuracy under GT-guided PGD-100 and near-zero clean accuracy support this interpretation (Section 5.5). The latter limits its use as a standalone defense.

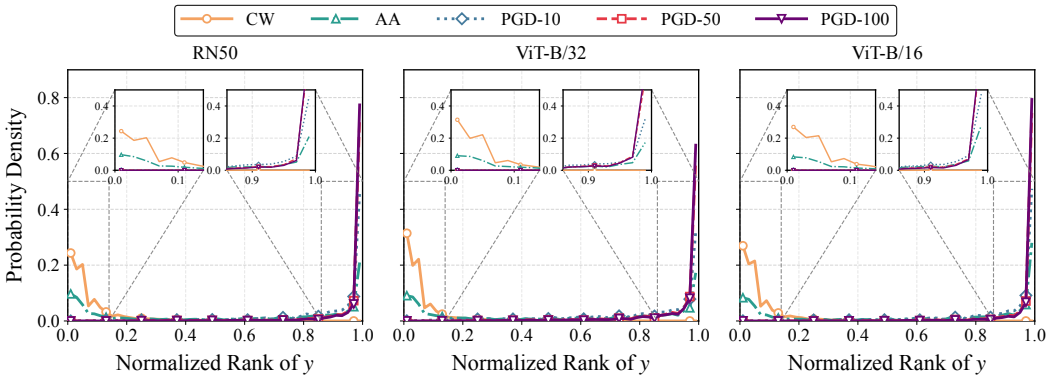


Figure 2: Comparison of the rank distribution of the ground-truth label on fine-grained datasets across multiple backbones under different adversarial attacks.

5 Experiments

We organize the experiments around four questions:

- **RQ1:** What label information does the attack leave in the feature?
- **RQ2:** Can transformed views recover the reverse attack direction?
- **RQ3:** How do margin gains depend on the label used by the attack?
- **RQ4:** What does direction-only prediction reveal about the recovered direction?

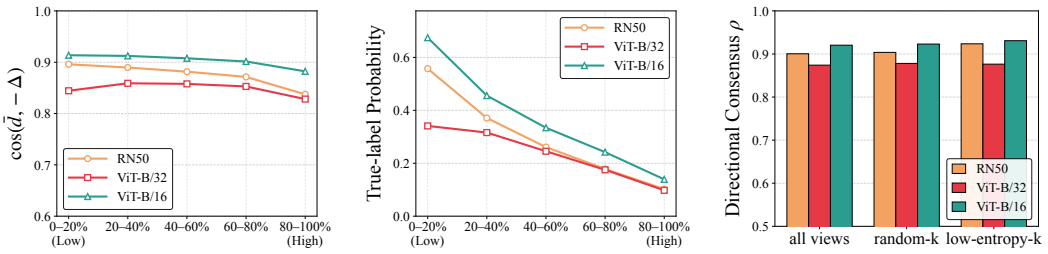
5.1 Experimental Setup

Benchmarks and backbones. Following DBD (Liu et al., 2026), we evaluate 15 datasets: five fine-grained datasets (Pets (Parkhi et al., 2012), Cars (Krause et al., 2013), Flower102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), and Aircraft (Maji et al., 2013)); five general datasets (Caltech101 (Fei-Fei et al., 2007), SUN397 (Xiao et al., 2010), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), and UCF101 (Soomro et al., 2012)); ImageNet (Deng et al., 2009); and four variants (ImageNet-A (Hendrycks et al., 2021b), ImageNetV2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a), and ImageNet-Sketch (Wang et al., 2019)). We use CLIP (Radford et al., 2021) with RN50 (He et al., 2016), ViT-B/32, and ViT-B/16 (Dosovitskiy et al., 2021), keeping DBD's transformations and hyperparameters. Per-dataset results across attacks are provided in Appendix B.

Evaluation details. Unless specified, we generate adversarial examples with GT-guided, non-adaptive white-box attacks on zero-shot CLIP and then apply the defense to the attacked inputs. PGD maximizes cross-entropy for the supplied attack label while keeping the image encoder and text embeddings fixed. PGD-10 and PGD-100 (Madry et al., 2018) use an ℓ_∞ budget of $\varepsilon = 4/255$ with 10 and 100 steps, respectively. Label-ablation experiments vary the attack label while holding the defense fixed. Unless noted otherwise, statistics aggregate all evaluated samples, and results are rounded to one decimal place. Adaptive attacks optimize the final defended loss through the full pipeline and are reported in Appendix C.5.

5.2 Label Information in Attacked Features (RQ1)

Label ranks and recovery. We examine true-label ranks to identify label information left by CW (Carlini & Wagner, 2017), AutoAttack (Croce & Hein, 2020), and PGD. Figure 2 shows that more PGD steps push the true label toward the lowest logit. CW favors rank 2, since only one competing class needs to overtake the true class. These results show that misclassification under the usual maximum-logit rule can coexist with a predictable position of the true label. We measure true-label coverage P at rank 2 for CW and at the lowest logit for other attacks. For non-CW attacks, P equals the accuracy of $h_{\min}(\hat{\mathbf{x}}) = \arg\min_c s_c(\hat{\mathbf{x}})$. On ViT-B/16 fine-grained datasets, P rises from 31.6% under PGD-10 to 66.9% under PGD-100. Full results appear in Table 5 in Appendix B.1.



(a) Defense Direction Alignment (b) True-Label Probability (c) Directional Consensus

Figure 3: Direction recovery under PGD-100 on fine-grained datasets: (a) reverse-direction alignment, (b) true-label probability, and (c) directional consensus.

Table 1: View-selection analysis with ViT/B-16 across datasets. ρ and cos (in %) quantify directional consensus and alignment, and Robust measures defended accuracy.

		CW			AutoAttack			PGD-100		
Dataset	View Selection	ρ	cos	Robust	ρ	cos	Robust	ρ	cos	Robust
Fine-grained	All views	72.0	49.0	52.6	83.1	61.6	69.0	95.2	94.5	91.9
	Random-k	73.0	48.4	53.5	83.4	61.5	69.0	95.4	94.4	91.8
	Low-entropy-k	76.0	50.5	56.8	85.1	61.5	69.1	96.0	94.6	91.9
General	All views	73.7	50.5	53.2	83.1	59.8	68.2	94.8	93.0	95.7
	Random-k	74.6	49.9	53.6	83.8	59.6	68.2	95.0	92.9	95.7
	Low-entropy-k	78.3	50.0	56.3	85.6	59.5	67.2	95.9	93.0	95.7
ImageNet	All views	73.2	52.7	57.2	83.4	63.8	71.9	95.1	95.0	97.6
	Random-k	74.2	52.2	58.1	83.9	63.7	71.9	95.3	94.8	97.6
	Low-entropy-k	77.2	53.8	61.6	85.5	63.7	72.0	95.9	95.1	97.7
ImageNet-Variants	All views	70.7	42.2	47.3	80.1	53.6	62.4	94.4	93.7	97.7
	Random-k	71.7	41.7	47.8	80.6	53.5	62.4	94.6	93.5	97.7
	Low-entropy-k	74.7	43.6	51.7	82.3	53.7	63.0	95.3	93.8	97.7

5.3 Direction Recovery from Transformed Views (RQ2)

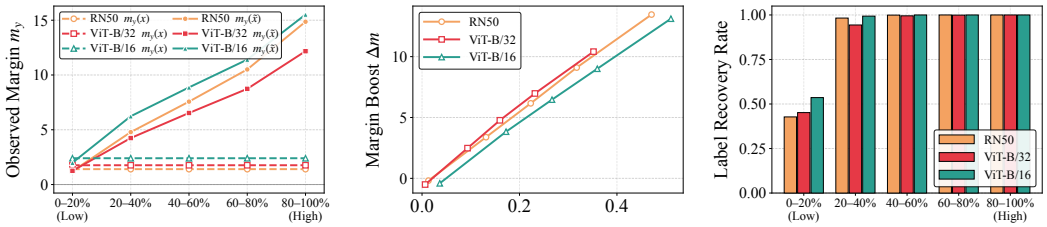
Entropy and direction. We group views by entropy to examine its relation to direction recovery. Under PGD-100, Figure 3 shows higher true-label probability and reverse-direction alignment for low-entropy than highest-entropy views. Consensus changes less across view-selection methods. Alignment measures recovery of the attack shift, whereas ρ measures agreement among views; the two capture different aspects of the recovered direction.

View selection. To test whether entropy filtering improves accuracy, we compare all views, random-k, and low-entropy-k. Random-k controls for reducing the number of views without using entropy. On general datasets, Table 1 shows that low-entropy selection raises CW accuracy from 53.2% to 56.3%, but lowers AutoAttack accuracy from 68.2% to 67.2%, despite higher ρ in both cases. Under PGD-100, accuracies remain similar across the three strategies, suggesting that useful recovery information is already available without entropy filtering. The effect of entropy filtering therefore depends on the attack; high consensus alone does not guarantee correct label recovery. Full results appear in Appendix B.2.

Recovery condition. We measure r to check the sufficient condition in Proposition 3.2. On ViT-B/16 fine-grained datasets, median r is 0.4 for PGD-10 and 0.3 for PGD-100; $r < 1$ holds for 99.5% and 100.0% of samples, respectively. The corresponding alignments are 91.4% and 94.6%. CW and AutoAttack have larger residual ratios and lower alignment. See Table 14 in Appendix C.1.

5.4 Margin Gain and Label Dependence (RQ3)

Margin gain and attack labels. We group PGD-100 samples by predicted gain G to examine its relation to margin change and label recovery. Figure 4 shows that higher-gain groups have larger

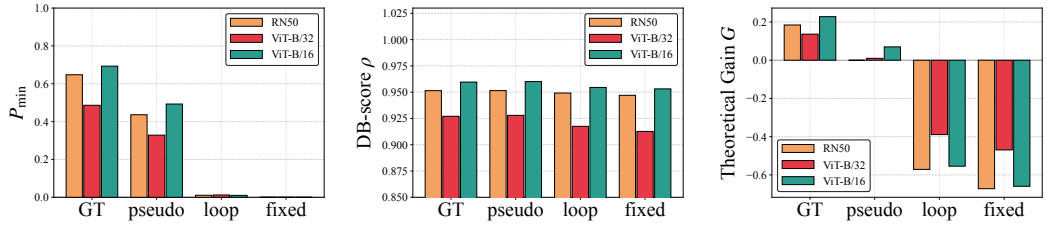


(a) Feature Margins (b) Observed Margin Gain (c) Label Recovery

Figure 4: Margins and label recovery on fine-grained datasets under PGD-100, grouped by predicted gain G: (a) clean and reconstructed margins, (b) margin change, and (c) label recovery.

Table 2: Mean predicted gain G, margin change Δm, and DBD accuracy under PGD-100. FG, Gen, IN, and IN-Var denote fine-grained, general, ImageNet, and its variants.

Metric	RN50				ViT-B/16				ViT-B/32			
	FG	Gen	IN	IN-Var	FG	Gen	IN	IN-Var	FG	Gen	IN	IN-Var
G	0.3	0.2	0.2	0.2	0.3	0.2	0.3	0.2	0.2	0.2	0.2	0.2
Δm	7.7	5.4	7.4	7.9	7.5	5.8	7.4	7.7	5.9	4.6	6.0	6.5
Robust	89.5	91.9	96.4	95.8	91.9	95.7	97.7	97.7	89.2	93.3	95.2	94.4



(a) True-Label Coverage (b) Directional Consensus (c) Predicted Gain

Figure 5: Effect of attack labels under PGD-100: (a) true-label minimum-logit coverage, (b) directional consensus, and (c) predicted gain.

reconstructed margins and better recovery; the lowest-gain group can still fall below the clean margin. This supports the analysis: reconstruction along directions with larger predicted gain tends to improve the true-class margin. To test whether recovery depends on the attack label, we keep the defense fixed and use GT, pseudo (clean CLIP prediction), loop (cyclic class permutation), or fixed labels. Figure 5 shows that loop and fixed labels reduce true-label minimum-logit coverage and make G negative despite high ρ. View agreement alone does not ensure true-label recovery.

Margin gains and defended accuracy. We compare mean margin changes and defended accuracy to quantify the effects of reconstruction and attack-label choice. Table 2 reports positive mean margin changes of 4.6–7.9 across all dataset groups and backbones under GT-guided PGD-100. Table 3 shows that DBD accuracy falls from 89.2–97.7% with GT labels to 41.6–69.9% with pseudo labels and at most 5.9% with loop or fixed labels. With the defense unchanged, this drop indicates that the observed recovery depends on the label used to generate the attack. To check the margin analysis, we measure M and B and estimate predicted gain and observed margin change from separate view subsets, as reported in Tables 16 and 17 of Appendix C.1. To further test label dependence, we vary the GT-label proportion and attack budget and compare clean-correct and clean-incorrect samples; see Tables 18 and 19 of Appendix C.2 and Table 20 of Appendix C.3.

5.5 Direction-Only Prediction (RQ4)

Direction-only prediction. We compare D-Pred and DBD using the same direction estimate. Removing the attacked-feature term tests whether the direction alone carries predictive label information.

Table 3: DBD accuracy (%) under PGD-100 using GT, pseudo, loop, or fixed attack labels.

Method	Label Type	RN50				ViT-B/16				ViT-B/32			
		FG	Gen	IN	IN-Var	FG	Gen	IN	IN-Var	FG	Gen	IN	IN-Var
CLIP (clean)		60.7	59.5	61.5	43.5	68.0	65.2	69.7	60.1	62.8	64.7	64.4	49.8
DBD	GT	89.5	92.1	96.3	95.7	91.9	96.9	97.7	97.7	89.2	93.2	95.4	94.3
	pseudo	59.4	58.8	60.8	41.6	67.5	66.9	69.9	60.5	61.9	64.1	63.8	49.4
	loop	2.3	1.8	1.6	1.9	2.6	1.4	2.2	1.6	4.6	4.4	5.2	4.4
	fixed	0.2	2.0	0.1	0.3	0.3	1.4	0.3	0.5	1.4	5.9	2.0	2.5

Table 4: Clean and PGD-100 accuracy (%). **D-Pred** classifies using only the recovered direction $\bar{\mathbf{d}}$. Best results within each backbone and column are in bold.

	Method	Fine-grained		General		ImageNet		ImageNet-Variants		Avg.	
		Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust	Clean	Robust
RN50	CLIP	60.7	0.0	**59.5**	0.0	61.5	0.0	43.5	0.0	**56.4**	0.0
	DBD	**61.7**	89.5	58.4	91.9	**63.1**	**96.4**	**45.0**	95.8	56.2	93.4
	D-Pred	1.6	**91.9**	2.9	**95.3**	0.6	95.6	1.8	**96.1**	2.0	**94.7**
ViT-B/32	CLIP	62.8	0.0	**64.7**	0.0	64.4	0.0	49.8	0.0	60.1	0.0
	DBD	**64.3**	89.2	63.9	93.3	**66.3**	**95.2**	**52.2**	94.4	**61.1**	93.0
	D-Pred	1.7	**91.2**	3.7	**96.6**	0.5	94.3	2.0	**95.0**	2.4	**94.3**
ViT-B/16	CLIP	68.2	0.0	**67.6**	0.0	69.6	0.0	60.1	0.0	65.9	0.0
	DBD	**69.3**	91.9	66.3	95.7	**71.1**	**97.7**	**61.9**	**97.7**	**66.4**	95.8
	D-Pred	2.5	**93.8**	3.6	**97.3**	0.7	97.2	2.4	97.2	2.7	**96.4**

Table 4 shows that D-Pred, defined in Eq. 19, achieves 94.3–96.4% average accuracy under GT-guided PGD-100, exceeding DBD for each backbone despite clean accuracies of only 2.0–2.7%. This contrast supports its use as a diagnostic of attack-induced label information. DBD remains higher on ImageNet for all three backbones and on ImageNet variants for ViT-B/16. These exceptions show that the attacked feature can still aid prediction despite D-Pred's higher average accuracy. Per-dataset comparisons appear in Appendix B.5.

Other defenses. We append DBD to R-TPT, TAPT, and TTC to test its gains across defenses. Under GT-guided PGD-100, their mean accuracies rise from 56.9%, 49.5%, and 68.9% to 89.2%, 91.4%, and 91.1%, respectively, with improvements on all ten datasets; see Table 21 in Appendix C.4.

Adaptive attacks. We optimize the defended loss through transformations, entropy selection, direction estimation, and reconstruction. Relative to attacks on vanilla CLIP, adaptive PGD-100 with EOT and BPDA (Athalye et al., 2018) reduces mean accuracy from 93.4% to 89.7% for DBD and from 91.8% to 89.1% for D-Pred. These drops show that the attack target affects recovery; see Table 23 in Appendix C.5.

6 Societal Impact, Limitation and Conclusion

Societal Impact and Limitation. This work improves understanding of adversarial defenses and may support safer AI systems, but carries dual-use risk by inspiring stronger attacks. It also shows that DBD's test-time robustification can be "parasitic robustness": a reverse-engineering mechanism whose behavior changes with the attack-label assignment, calling for label-assignment controls.

Conclusion. We deconstruct DBD and reveal a counterintuitive paradox: adversarial attacks can leave clear forensic trails by encoding the ground-truth label as the global minimum logit. Random image transformations move features toward the GT anchor, and predictive-entropy filtering isolates an aligned reverse trajectory that converts the attacker's perturbation into margin gain. Experiments with ground-truth, pseudo, and surrogate attack labels quantify this dependence.

AI USE STATEMENT

Generative AI tools were used to assist with writing and revising the manuscript. The authors carefully reviewed and verified all AI-assisted content and take full responsibility for the accuracy and integrity of the final manuscript.

REPRODUCIBILITY STATEMENT

We describe the datasets, evaluation protocols, baselines, and main experimental setup in Section 5. Detailed algorithmic and implementation information is provided in the appendices, together with theoretical proofs, comprehensive dataset-level results, and additional control experiments and evaluations.

REFERENCES

Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, Proceedings of Machine Learning Research, pp. 274–283. PMLR, 2018.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *ECCV (6)*, Lecture Notes in Computer Science, pp. 446–461. Springer, 2014.

Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57. IEEE Computer Society, 2017.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pp. 3606–3613. IEEE Computer Society, 2014.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, Proceedings of Machine Learning Research, pp. 2206–2216. PMLR, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. IEEE Computer Society, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021.

Li Fei-Fei, Robert Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, 2007.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR (Poster)*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778. IEEE Computer Society, 2016.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7):2217–2226, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pp. 8320–8329. IEEE, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pp. 15262–15271. Computer Vision Foundation / IEEE, 2021b.

Chengze Jiang, Minjing Dong, Xinli Shi, and Jie Gui. Diversifying counterattacks: Orthogonal exploration for robust CLIP inference. *CoRR*, abs/2511.09064, 2025.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, pp. 554–561. IEEE Computer Society, 2013.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *ICLR (Poster)*. OpenReview.net, 2017.

Jiaxiang Liu, Jiawei Du, Xiao Liu, Prayag Tiwari, and Mingkun Xu. Self-calibrated consistency can fight back for adversarial robustness in vision-language models. *CoRR*, abs/2510.22785, 2025.

Liangsheng Liu, Si Chen, Jiamin Wu, Weiwei Feng, Zhixin Cheng, Xiaotian Yin, Wenfei Yang, and Tianzhu Zhang. Adversarial attacks already tell the answer: Directional bias-guided test-time defense for vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UqC2oFRRyc.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR (Poster)*. OpenReview.net, 2018.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *ICLR*. OpenReview.net, 2023.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pp. 722–729. IEEE Computer Society, 2008.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, pp. 3498–3505. IEEE Computer Society, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, Proceedings of Machine Learning Research, pp. 8748–8763. PMLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, Proceedings of Machine Learning Research, pp. 5389–5400. PMLR, 2019.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *ICML*, Proceedings of Machine Learning Research, pp. 43685–43704. PMLR / OpenReview.net, 2024.

Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-TPT: improving adversarial robustness of vision-language models through test-time prompt tuning. In *CVPR*, pp. 29958–29967. Computer Vision Foundation / IEEE, 2025.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR (Poster)*, 2014.

Baoshun Tong, Hanjiang Lai, Yan Pan, and Jian Yin. On the zero-shot adversarial robustness of vision-language models: A truly zero-shot and training-free approach. In *CVPR*, pp. 19921–19930. Computer Vision Foundation / IEEE, 2025.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *NeurIPS*, 2020.

Chenguang Wang, Ruoxi Jia, Xin Liu, and Dawn Song. Benchmarking zero-shot robustness of multimodal foundation models: A pilot study. *CoRR*, abs/2403.10499, 2024a.

Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, pp. 10506–10518, 2019.

Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *CVPR*, pp. 24502–24511. IEEE, 2024b.

Xin Wang, Kai Chen, Jiaming Zhang, Jingjing Chen, and Xingjun Ma. TAPT: test-time adversarial prompt tuning for robust inference in vision-language models. In *CVPR*, pp. 19910–19920. Computer Vision Foundation / IEEE, 2025.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pp. 3485–3492. IEEE Computer Society, 2010.

Songlong Xing, Zhengyu Zhao, and Nicu Sebe. CLIP is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of CLIP. In *CVPR*, pp. 15172–15182. Computer Vision Foundation / IEEE, 2025.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark A. Hasegawa-Johnson, Yingzhen Li, and Chang D. Yoo. C-TPT: calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *ICLR*. OpenReview.net, 2024.

Jiacheng Zhang, Jinhao Li, Hanxun Huang, Sarah Monazam Erfani, Benjamin I. P. Rubinstein, and Feng Liu. Semantic-aware adversarial fine-tuning for CLIP. *Trans. Mach. Learn. Res.*, 2026, 2026a.

Ruize Zhang, Yu Li, Lei Li, Zhang Wan, Juan Cao, and Sheng Tang. Adversaries fight back: Revisiting test-time adversarial defenders of vision-language model under adaptive attacks, 2026b. URL https://openreview.net/forum?id=hZVSKjtaS5.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *ICML*, Proceedings of Machine Learning Research, pp. 42058–42080. PMLR, 2023.

Mo Zhou and Vishal M. Patel. On trace of pgd-like adversarial attacks. In *ICPR (3)*, Lecture Notes in Computer Science, pp. 81–97. Springer, 2024.

Yiwei Zhou, Xiaobo Xia, Zhiwei Lin, Bo Han, and Tongliang Liu. Few-shot adversarial prompt learning on vision-language models. In *NeurIPS*, 2024.

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 Problem Setup and Analysis **3**
3.1 Preliminaries . . . 3
3.2 Overview of DBD . . . 4
3.3 Theoretical Analyses of DBD under GT-guided Attack . . . 4

4 D-Pred: A Direction-Only Method **5**

5 Experiments **6**
5.1 Experimental Setup . . . 6
5.2 Label Information in Attacked Features (RQ1) . . . 6
5.3 Direction Recovery from Transformed Views (RQ2) . . . 7
5.4 Margin Gain and Label Dependence (RQ3) . . . 7
5.5 Direction-Only Prediction (RQ4) . . . 8

6 Societal Impact, Limitation and Conclusion **9**

A Proofs for Section 3.3 **14**
A.1 Gradient Structure of the CW Loss . . . 14
A.2 Proof of Proposition 3.1 . . . 14
A.3 Proof of Proposition 3.2 . . . 14
A.4 Proof of Proposition 3.3 . . . 15

B Comprehensive Dataset-Level Experimental Results **15**
B.1 Detailed Results for RQ1: Ground-Truth Label Leakage . . . 15
B.2 Detailed Results for RQ2: Efficacy of Entropy Filtering . . . 19
B.3 Detailed Results for RQ3: Directional Extrapolation and Margin Gain . . . 23
B.4 Detailed Results for RQ3: Surrogate-Label Ablation . . . 29
B.5 Detailed Results for RQ4: CLIP vs. DBD vs. **D-Pred** . . . 42

C Label-Reliance Controls and Additional Evaluations **45**
C.1 Empirical Checks of Theoretical Quantities . . . 45
C.2 Label-Purity and Conditional Controls . . . 45
C.3 Perturbation-Budget Controls . . . 46
C.4 Transfer to Other Test-Time Defenses . . . 46
C.5 Complete-Pipeline Adaptive Evaluation . . . 47

A PROOFS FOR SECTION 3.3

A.1 GRADIENT STRUCTURE OF THE CW LOSS

We verify the claim in Section 3.3 that CW suppresses s_y relative to its closest competitor. Consider a feature $\mathbf{x}$ for which y is the top-logit class, and let $c^\star = \arg\max_{c \neq y} s_c(\mathbf{x})$ denote the runner-up.

Proposition A.1 (CW loss suppresses s_y unconditionally)**.** *Let $\mathcal{L}_{\mathrm{CW}}(\mathbf{x}, y) = s_y(\mathbf{x}) - s_{c^\star}(\mathbf{x})$. For any $\mathbf{x}$ at which $c^\star$ is unique, the feature-space gradient equals*

$$\nabla_{\mathbf{x}} \mathcal{L}_{\mathrm{CW}} = \mathbf{w}_y - \mathbf{w}_{c^\star}, \tag{20}$$

and the attacker's descent direction $-\nabla \mathcal{L}_{\mathrm{CW}}$ satisfies

$$\mathbf{w}_y^\top (-\nabla_{\mathbf{x}} \mathcal{L}_{\mathrm{CW}}) = \mathbf{w}_y^\top \mathbf{w}_{c^\star} - 1 < 0. \tag{21}$$

Proof. Eq. equation 20 follows by linearity of $s_c(\mathbf{x}) = \mathbf{w}_c^\top \mathbf{x}$, treating $c^\star$ as a fixed index in a neighborhood where the argmax is stable. Eq. equation 21 then uses $\mathbf{w}_y^\top \mathbf{w}_y = 1$ and $\mathbf{w}_y^\top \mathbf{w}_{c^\star} < 1$ since $\mathbf{w}_{c^\star} \neq \mathbf{w}_y$. □

Thus CW reduces s_y through the runner-up direction; whether this is enough to make y the minimum-logit class is measured empirically in RQ1.

A.2 PROOF OF PROPOSITION 3.1

Proof of Proposition 3.1. For the frozen CLIP logits used by the cross-entropy attack,

$$\mathcal{L}_{\mathrm{CE}}(\mathbf{z}, y) = -\frac{s_y(\mathbf{z})}{\tau} + \log \sum_{c=1}^{C} \exp\left(\frac{s_c(\mathbf{z})}{\tau}\right), \qquad \nabla_{\mathbf{z}} \mathcal{L}_{\mathrm{CE}} = \frac{\boldsymbol{\mu}(\mathbf{z}) - \mathbf{w}_y}{\tau}.$$

Distinct unit text embeddings satisfy $\mathbf{w}_y^\top \mathbf{w}_c < 1$ for $c \neq y$; positive softmax weights imply:

$$\mathbf{w}_y^\top \boldsymbol{\mu}(\mathbf{z}) = \pi_y + \sum_{c \neq y} \pi_c (\mathbf{w}_y^\top \mathbf{w}_c) < \pi_y + \sum_{c \neq y} \pi_c = 1.$$

Thus

$$\mathbf{w}_y^\top \nabla_{\mathbf{z}} \mathcal{L}_{\mathrm{CE}} = \frac{1}{\tau} (\mathbf{w}_y^\top \boldsymbol{\mu}(\mathbf{z}) - 1) < 0.$$

This gives $\mathbf{w}_y^\top \nabla_{\mathbf{z}} \mathcal{L}_{\mathrm{CE}} < 0$.

For $c \neq y$, let $\mathbf{q}_c = \mathbf{w}_y - \mathbf{w}_c$. Then

$$\mathbf{q}_c^\top \nabla_{\mathbf{z}} \mathcal{L}_{\mathrm{CE}} = \frac{1}{\tau} \left[-\pi_c(\mathbf{z}) \|\mathbf{q}_c\|^2 + \sum_{j \notin \{y, c\}} \pi_j(\mathbf{z}) \mathbf{q}_c^\top (\mathbf{w}_j - \mathbf{w}_y) \right] = \frac{1}{\tau} \left[-\pi_c(\mathbf{z}) \|\mathbf{q}_c\|^2 + E_c(\mathbf{z}) \right].$$

Since $\tau > 0$, this quantity is negative exactly when $\pi_c(\mathbf{z}) \|\mathbf{q}_c\|^2 > E_c(\mathbf{z})$. □

A.3 PROOF OF PROPOSITION 3.2

Proof of Proposition 3.2. By the definitions of $\bar{\mathbf{d}}$ and $\bar{\boldsymbol{\xi}}$,

$$\bar{\mathbf{d}} = \frac{1}{k} \sum_{i=1}^{k} (\hat{\mathbf{x}}_i - \hat{\mathbf{x}}) = \frac{1}{k} \sum_{i=1}^{k} \left[(\hat{\mathbf{x}}_i - \mathbf{x}) - (\hat{\mathbf{x}} - \mathbf{x}) \right] = \bar{\boldsymbol{\xi}} - \boldsymbol{\Delta},$$

which is Eq. equation 12. Let $\mathbf{u} = -\boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|$ and $\bar{\boldsymbol{\xi}}/\|\boldsymbol{\Delta}\| = a\mathbf{u} + \mathbf{v}$, where $\mathbf{v} \perp \mathbf{u}$. $a^2 + \|\mathbf{v}\|^2 = r^2$ and

$$\frac{\bar{\mathbf{d}}}{\|\boldsymbol{\Delta}\|} = (1 + a)\mathbf{u} + \mathbf{v}.$$

Hence its alignment with $\mathbf{u}$ is

$$\frac{1 + a}{\sqrt{(1 + a)^2 + \|\mathbf{v}\|^2}} = \frac{1 + a}{\sqrt{1 + 2a + r^2}}.$$

For $r < 1$, the minimum over $a \in [-r, r]$ is $\sqrt{1 - r^2}$ at $a = -r^2$, proving Eq. equation 13. □

A.4 Proof of Proposition 3.3

Proof of Proposition 3.3. For each $c \neq y$, Eq. equation 16 gives

$$\mathbf{q}_c^\top \tilde{\mathbf{x}} - m_y(\mathbf{x}) = \gamma_c + \big(\beta(\rho) - 1\big) A_c + \beta(\rho) \mathbf{q}_c^\top \bar{\boldsymbol{\xi}}.$$

By Cauchy–Schwarz,

$$\mathbf{q}_c^\top \bar{\boldsymbol{\xi}} \geq -\|\mathbf{q}_c\| \, \|\bar{\boldsymbol{\xi}}\|.$$

Thus every pairwise margin at $\tilde{\mathbf{x}}$ is larger than $m_y(\mathbf{x}) + B$. Taking the minimum over $c \neq y$ gives $m_y(\tilde{\mathbf{x}}) - m_y(\mathbf{x}) \geq B > 0$. □

B Comprehensive Dataset-Level Experimental Results

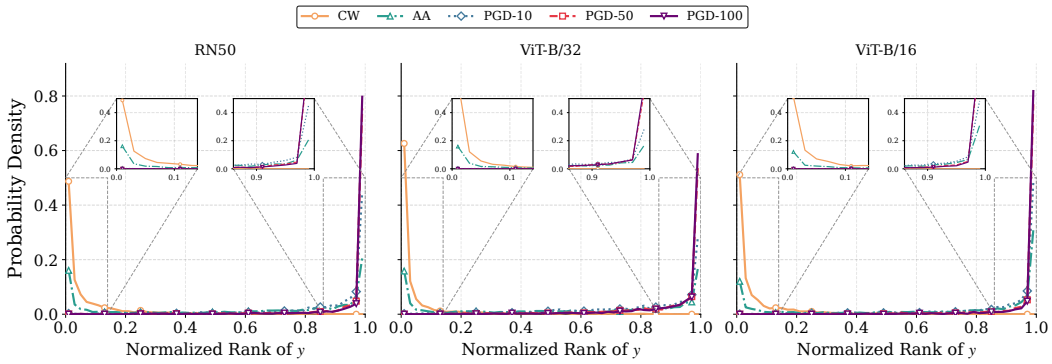
Benchmarks and backbones. We follow the DBD (Liu et al., 2026) benchmark suite, including fine-grained datasets (Pets (Parkhi et al., 2012), Cars (Krause et al., 2013), Flower102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), and Aircraft (Maji et al., 2013)), general recognition datasets (Caltech101 (Fei-Fei et al., 2007), SUN397 (Xiao et al., 2010), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), and UCF101 (Soomro et al., 2012)), ImageNet (Deng et al., 2009), and shifted ImageNet variants (ImageNet-A (Hendrycks et al., 2021b), ImageNetV2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a), and ImageNet-Sketch (Wang et al., 2019)). We use RN50, ViT-B/32, and ViT-B/16 CLIP backbones (He et al., 2016; Dosovitskiy et al., 2021; Liu et al., 2026), with the same transformations and hyperparameters as DBD. Unless noted, experiments use the GT-guided setting from our analysis.

The main text reports group-level summaries; this appendix provides the full per-dataset measurements across all 15 datasets and three CLIP backbones. Figures are drawn from four representative datasets (Flower102, Caltech101, ImageNet, ImageNet-R) for readability, while tables cover every dataset–backbone–attack combination. Unless noted otherwise, tables report all-sample aggregates.

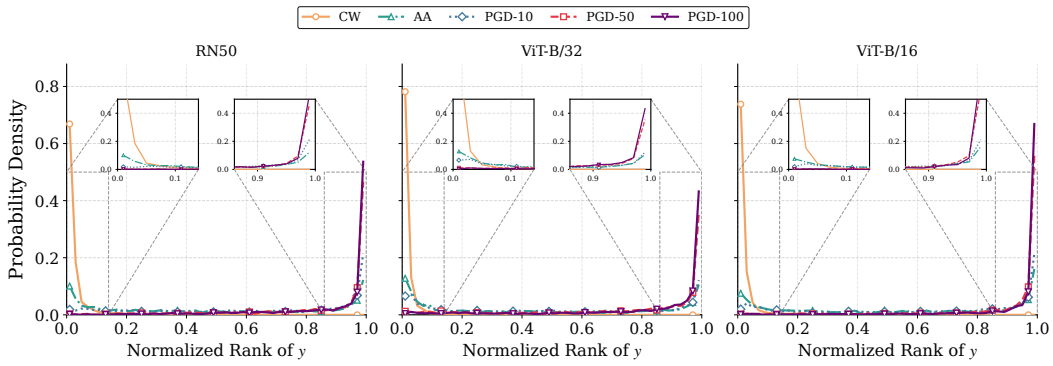
B.1 Detailed Results for RQ1: Ground-Truth Label Leakage

B.1.1 Rank Distribution Visualizations

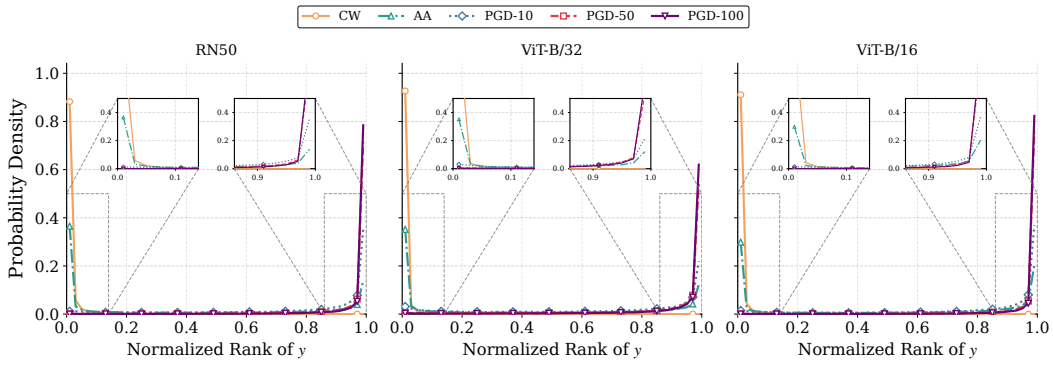
Figure 6 shows the post-attack rank of the ground-truth label across four datasets, with one column per backbone (RN50, ViT-B/32, ViT-B/16). The horizontal axis is normalized rank, so rightward shift indicates the true label is being pushed below an increasing fraction of candidate classes. PGD progressively concentrates probability mass in the far-right tail as the iteration budget grows, reflecting the minimum-logit pattern predicted by the theory. CW, in contrast, primarily exposes a rank-2 pattern consistent with its margin-based objective, and does not reach the same extreme.



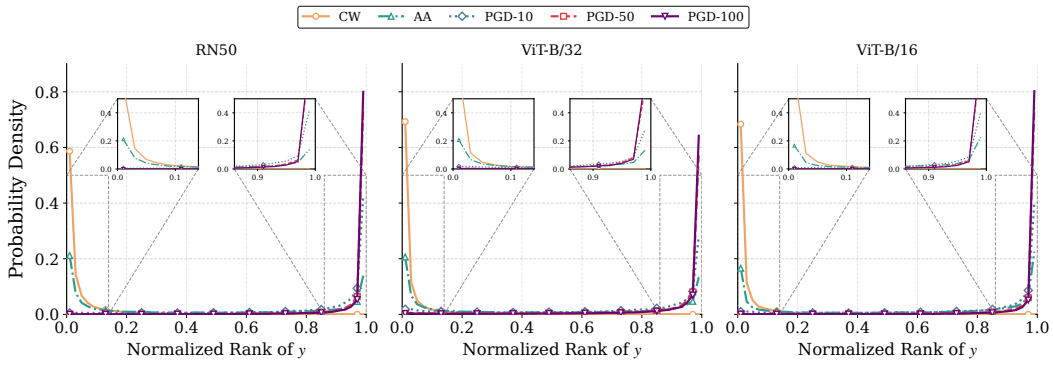
(a) Flower102



(b) Caltech101



(c) ImageNet



(d) ImageNet-R

Figure 6: Representative RQ1 rank distributions for Flower102, Caltech101, ImageNet, and ImageNet-R.

B.1.2 Quantitative Leakage Metrics

Table 5 reports group-level leakage statistics, and Table 6 reports per-dataset results. The P column denotes the probability of the attack-specific leakage event: for CW, P measures $\Pr(\text{rank} = 2)$, *i.e.*, the probability that exactly one competing class overtakes the true label; for non-CW attacks, P measures $\Pr(y = \arg\min_c s_c(\hat{\mathbf{x}}))$, the minimum-logit condition. Rank is the raw mean post-attack rank, with rank 1 denoting the highest logit. The logit shifts are $S_y = s_y(\hat{\mathbf{x}}) - s_y(\mathbf{x})$ and $S_{\neg y} = \frac{1}{C-1}\sum_{c \neq y}(s_c(\hat{\mathbf{x}}) - s_c(\mathbf{x}))$; the tables report their sample means. For non-CW attacks, P is also the accuracy of $h_{\min}(\hat{\mathbf{x}}) = \arg\min_{c \in [C]} s_c(\hat{\mathbf{x}})$. The PGD rows exhibit the strongest monotone trend: on ViT-B/16 Flower102, PGD-100 reaches $P = 68.8$ with mean rank 99.0 and a substantially larger negative shift in the true-label logit than in the non-true logits.

Table 5: The label leakage phenomenon under adversarial attacks across backbones and dataset groups. P (%) denotes attack-specific leakage coverage: rank 2 for CW, and minimum logit for other attacks. S_y and $S_{\neg y}$ represent the average logit changes for the ground-truth class and other classes.

	Attack	Fine-grained			General			ImageNet			ImageNet-Variants		
		P	S_y	$S_{\neg y}$	P	S_y	$S_{\neg y}$	P	S_y	$S_{\neg y}$	P	S_y	$S_{\neg y}$
RN50	CW	35.1	-2.7	0.0	42.1	-1.9	0.3	34.8	-2.9	-0.1	32.6	-1.8	0.0
	AA	13.9	-12.8	-1.3	10.6	-8.2	-0.7	3.9	-12.0	-0.9	4.6	-7.4	-0.6
	PGD-10	30.7	-22.7	-3.5	29.5	-13.9	-1.9	12.3	-20.8	-2.1	22.7	-18.1	-2.2
	PGD-50	58.0	-32.4	-6.1	53.4	-20.0	-3.7	41.0	-29.8	-3.6	49.5	-25.9	-3.7
	PGD-100	64.2	-35.1	-6.9	59.7	-21.8	-4.2	49.2	-32.3	-4.0	56.8	-28.1	-4.1
ViT-B/16	CW	39.5	-3.3	0.0	44.3	-3.0	-0.2	42.5	-3.3	-0.1	42.8	-2.8	0.1
	AA	18.9	-17.8	-3.0	19.2	-11.3	-1.8	5.8	-15.8	-1.8	9.4	-12.1	-1.3
	PGD-10	31.6	-26.9	-5.3	34.7	-16.7	-3.4	12.5	-23.2	-3.0	23.1	-20.8	-2.8
	PGD-50	60.5	-37.1	-7.9	63.2	-23.9	-4.9	47.1	-32.9	-4.1	53.8	-29.5	-4.0
	PGD-100	66.9	-39.8	-8.5	69.8	-26.0	-5.2	56.5	-35.6	-4.3	61.2	-32.0	-4.2
ViT-B/32	CW	46.6	-2.1	0.3	47.5	-2.1	0.1	45.5	-2.4	0.2	41.6	-1.7	0.2
	AA	10.4	-13.1	-1.6	13.4	-9.2	-1.2	2.9	-12.6	-1.2	4.7	-8.5	-0.8
	PGD-10	19.2	-20.4	-3.0	26.3	-13.4	-2.4	5.7	-18.9	-2.0	14.7	-16.7	-2.1
	PGD-50	41.9	-27.5	-4.8	47.1	-18.8	-3.6	24.7	-25.8	-3.0	36.0	-23.0	-3.2
	PGD-100	48.1	-29.5	-5.3	53.2	-20.2	-3.9	31.8	-27.7	-3.2	42.4	-24.7	-3.4

Table 6: Dataset-level RQ1 leakage metrics aggregated from all currently available raw data (all-sample scope). P (%) denotes the probability of the ground-truth label reaching the attack-specific leakage state: rank 2 for CW, and minimum logit for all other attacks.

Backbone: RN50 Attack: CW

Dataset	P	Rank	Δ_y	$\Delta_{\neg y}$
Aircraft	15.4	12.6	-0.5	0.1
Cars	34.7	6.9	-2.2	0.0
Flower102	35.8	10.0	-2.9	0.0
Food101	34.5	9.2	-4.4	-0.3
Pets	59.1	3.0	-3.3	0.0
Caltech101	66.9	3.8	-3.2	0.9
DTD	44.2	5.4	-1.2	0.1
EuroSAT	39.1	3.9	-0.3	0.3
SUN397	41.0	7.6	-2.5	-0.1
UCF101	37.6	7.2	-2.0	0.1
ImageNet	34.8	12.6	-2.9	-0.1
ImageNet-A	15.9	16.9	-1.3	0.0
ImageNet-R	38.0	10.8	-2.6	0.0
ImageNet-Sketch	32.0	27.7	-0.8	0.0
ImageNetV2	32.4	15.2	-2.5	0.0

Backbone: RN50 Attack: PGD-10

Dataset	P	Rank	Δ_y	$\Delta_{\neg y}$
Aircraft	27.3	88.6	-18.4	-4.4
Cars	29.7	180.2	-31.7	-5.0
Flower102	27.5	89.1	-21.1	-2.8
Food101	50.5	94.0	-21.2	-2.2
Pets	18.5	29.1	-21.2	-2.9
Caltech101	15.8	64.8	-15.6	-0.8
DTD	28.0	37.7	-10.0	-1.7
EuroSAT	55.6	9.3	-11.2	-4.0
SUN397	15.2	311.5	-17.7	-2.2
UCF101	33.0	86.0	-15.0	-1.0
ImageNet	12.3	778.6	-20.8	-2.1
ImageNet-A	42.5	187.3	-18.6	-2.7
ImageNet-R	25.2	167.4	-17.8	-2.3
ImageNet-Sketch	9.0	729.9	-15.6	-1.9
ImageNetV2	14.3	797.9	-20.4	-2.0

Backbone: RN50 Attack: AutoAttack

Dataset	P	Rank	Δ_y	$\Delta_{\neg y}$
Aircraft	6.1	26.7	-3.5	-0.4
Cars	11.3	100.4	-16.6	-2.1
Flower102	11.8	61.2	-13.0	-1.2
Food101	29.2	69.5	-14.9	-1.1
Pets	10.8	23.5	-16.2	-1.8
Caltech101	8.3	49.8	-12.0	0.1
DTD	11.3	20.5	-5.0	-0.6
EuroSAT	12.6	5.8	-4.9	-1.9
SUN397	6.0	178.5	-10.6	-1.0
UCF101	14.8	51.2	-8.5	-0.1
ImageNet	3.9	416.9	-12.0	-0.9
ImageNet-A	6.6	52.9	-4.4	-0.4
ImageNet-R	7.0	89.5	-9.7	-0.8
ImageNet-Sketch	0.9	213.1	-5.1	-0.3
ImageNetV2	3.9	386.9	-10.6	-0.7

Backbone: RN50 Attack: PGD-50

Dataset	P	Rank	Δ_y	$\Delta_{\neg y}$
Aircraft	45.5	95.1	-25.4	-7.1
Cars	58.6	191.9	-42.1	-7.6
Flower102	59.1	96.7	-31.7	-5.8
Food101	78.3	99.1	-29.3	-4.1
Pets	48.6	33.6	-33.4	-6.0
Caltech101	38.0	85.2	-22.9	-2.6
DTD	48.3	41.6	-15.6	-3.9
EuroSAT	78.4	9.7	-16.5	-6.2
SUN397	44.5	372.5	-24.5	-3.6
UCF101	57.8	95.1	-20.5	-2.1
ImageNet	41.0	938.3	-29.8	-3.6
ImageNet-A	72.1	197.5	-26.6	-4.2
ImageNet-R	57.1	190.1	-25.7	-3.9
ImageNet-Sketch	25.7	874.5	-21.9	-3.2
ImageNetV2	43.1	943.7	-29.2	-3.5

Backbone: RN50 Attack: PGD-100

Dataset	P	Rank	Δ_y	$\Delta_{\neg y}$
Aircraft	51.0	96.5	-27.4	-7.7
Cars	65.1	193.2	-44.5	-8.2
Flower102	65.6	97.8	-34.8	-6.6
Food101	82.7	99.6	-31.5	-4.7
Pets	56.5	34.2	-37.3	-7.1
Caltech101	45.3	88.5	-25.0	-3.1
DTD	55.2	42.6	-17.5	-4.7
EuroSAT	82.3	9.8	-17.9	-6.8
SUN397	52.5	379.4	-26.4	-3.9
UCF101	63.4	96.7	-22.2	-2.4
ImageNet	49.2	956.9	-32.3	-4.0
ImageNet-A	79.1	198.4	-29.0	-4.6
ImageNet-R	65.3	193.0	-28.0	-4.4
ImageNet-Sketch	31.3	899.0	-23.7	-3.6
ImageNetV2	51.5	961.6	-31.8	-3.8

Backbone: ViT-B/16 Attack: CW

Dataset	P	Rank	Δ_y	$\Delta_{\neg y}$
Aircraft	21.1	8.7	-0.9	0.0
Cars	39.5	6.0	-3.1	-0.2
Flower102	37.8	8.7	-3.2	0.2
Food101	38.7	8.2	-5.1	0.1
Pets	63.4	2.9	-4.1	0.0
Caltech101	73.8	3.5	-4.4	0.4
DTD	37.4	5.9	-1.7	0.1
EuroSAT	33.8	4.4	-3.1	-1.2
SUN397	45.5	6.7	-2.8	-0.2
UCF101	44.2	6.6	-2.7	-0.1
ImageNet	42.5	10.6	-3.3	-0.1
ImageNet-A	23.9	17.2	-3.2	0.2
ImageNet-R	48.6	9.3	-3.6	0.1
ImageNet-Sketch	43.1	18.0	-1.3	0.1
ImageNetV2	38.1	13.5	-2.9	0.1

Backbone: ViT-B/16 Attack: AutoAttack

Dataset	P	Rank	Δ_y	$\Delta_{\neg y}$
Aircraft	8.5	30.6	-6.4	-1.1
Cars	20.4	118.6	-23.5	-4.3
Flower102	18.1	68.3	-18.2	-2.7
Food101	35.7	77.1	-19.5	-2.2
Pets	11.8	24.9	-21.4	-4.7
Caltech101	11.4	54.8	-15.7	-1.4
DTD	21.9	24.4	-7.2	-1.0
EuroSAT	35.5	6.9	-9.1	-3.7
SUN397	9.3	203.2	-13.3	-2.0
UCF101	17.7	56.9	-11.3	-1.0
ImageNet	5.8	518.1	-15.8	-1.8
ImageNet-A	17.6	96.6	-11.2	-1.2
ImageNet-R	12.0	115.6	-14.6	-1.6
ImageNet-Sketch	2.2	287.2	-8.2	-0.8
ImageNetV2	6.0	484.4	-14.3	-1.5

Backbone: ViT-B/16 Attack: PGD-10

Dataset	P	Rank	Δ_y	$\Delta_{\neg y}$
Aircraft	26.3	91.4	-23.6	-5.9
Cars	37.2	180.6	-37.1	-7.4
Flower102	31.1	89.8	-25.1	-4.3
Food101	47.5	92.3	-23.7	-3.1
Pets	15.6	28.8	-25.2	-5.7
Caltech101	15.6	63.5	-18.0	-2.1
DTD	35.2	39.2	-12.2	-2.5
EuroSAT	72.5	9.5	-16.0	-6.9
SUN397	17.1	312.4	-19.5	-3.2
UCF101	32.9	85.2	-17.8	-2.5
ImageNet	12.5	788.5	-23.2	-3.0
ImageNet-A	42.8	185.7	-21.8	-3.1
ImageNet-R	23.9	162.4	-20.1	-2.7
ImageNet-Sketch	10.0	712.4	-18.3	-2.5
ImageNetV2	15.4	812.2	-22.8	-2.8

Backbone: ViT-B/16 Attack: PGD-50

Dataset	P	Rank	Δ_y	$\Delta_{\neg y}$
Aircraft	47.1	96.2	-31.6	-8.9
Cars	66.0	191.6	-47.6	-9.7
Flower102	61.6	97.8	-35.0	-6.9
Food101	79.2	99.2	-31.9	-4.4
Pets	48.8	34.4	-39.2	-9.5
Caltech101	46.0	89.4	-26.6	-3.6
DTD	60.5	43.1	-18.0	-4.0
EuroSAT	89.5	9.9	-23.1	-9.1
SUN397	53.0	378.5	-27.2	-4.1
UCF101	66.8	96.5	-24.7	-3.4
ImageNet	47.1	947.6	-32.9	-4.1
ImageNet-A	76.3	197.5	-31.1	-4.2
ImageNet-R	58.2	189.2	-28.9	-4.1
ImageNet-Sketch	31.1	877.9	-25.8	-3.7
ImageNetV2	49.6	954.5	-32.4	-3.8

Backbone: ViT-B/16 Attack: PGD-100

Dataset	P	Rank	Δ_y	$\Delta_{\neg y}$
Aircraft	53.6	97.0	-33.9	-9.6
Cars	71.4	192.7	-50.0	-10.2
Flower102	68.8	99.0	-38.0	-7.5
Food101	84.2	99.8	-34.2	-4.7
Pets	56.6	35.2	-43.1	-10.4
Caltech101	57.1	92.7	-29.1	-3.8
DTD	65.1	43.7	-19.9	-4.5
EuroSAT	90.9	9.9	-25.3	-9.7
SUN397	62.1	385.0	-29.4	-4.4
UCF101	74.0	97.7	-26.5	-3.6
ImageNet	56.5	965.3	-35.6	-4.3
ImageNet-A	82.1	198.4	-33.7	-4.4
ImageNet-R	66.1	192.1	-31.4	-4.4
ImageNet-Sketch	37.7	904.1	-27.8	-3.9
ImageNetV2	58.9	969.7	-35.1	-4.0

Backbone: ViT-B/32 Attack: CW

Dataset	P	Rank	Δ_y	$\Delta_{\neg y}$
Aircraft	18.9	10.3	-0.4	0.1
Cars	44.1	4.8	-1.8	0.1
Flower102	46.9	7.4	-1.9	0.6
Food101	47.3	5.9	-3.5	0.3
Pets	70.4	2.6	-2.8	0.4
Caltech101	78.2	3.2	-3.2	0.9
DTD	41.4	5.5	-1.1	0.3
EuroSAT	32.8	4.3	-2.2	-0.8
SUN397	50.2	5.7	-2.2	0.0
UCF101	47.2	5.6	-1.9	0.3
ImageNet	45.5	8.7	-2.4	0.2
ImageNet-A	20.2	15.5	-1.5	0.2
ImageNet-R	48.8	8.3	-2.6	0.2
ImageNet-Sketch	40.4	22.3	-0.9	0.2
ImageNetV2	40.4	11.7	-2.0	0.2

Backbone: ViT-B/32 Attack: AutoAttack

Dataset	P	Rank	Δ_y	$\Delta_{\neg y}$
Aircraft	5.9	26.3	-3.9	-0.4
Cars	8.7	99.0	-16.5	-2.3
Flower102	8.0	57.6	-12.9	-1.4
Food101	20.4	67.7	-15.3	-1.4
Pets	9.3	22.6	-17.1	-2.4
Caltech101	7.4	45.3	-12.7	-0.5
DTD	15.3	22.2	-5.7	-0.7
EuroSAT	28.4	6.9	-7.3	-2.8
SUN397	4.3	168.1	-10.8	-1.4
UCF101	11.5	50.2	-9.5	-0.8
ImageNet	2.9	416.7	-12.6	-1.2
ImageNet-A	8.4	63.0	-6.0	-0.7
ImageNet-R	6.5	92.9	-10.9	-1.0
ImageNet-Sketch	1.1	219.5	-6.0	-0.4
ImageNetV2	2.9	386.4	-11.1	-1.0

Backbone: ViT-B/32 Attack: PGD-10

Dataset	P	Rank	Δ_y	$\Delta_{\neg y}$
Aircraft	24.2	84.8	-17.6	-3.6
Cars	18.3	163.9	-27.6	-4.4
Flower102	14.9	80.5	-18.2	-2.2
Food101	27.9	85.7	-18.9	-2.0
Pets	10.6	26.2	-19.8	-3.0
Caltech101	9.1	50.5	-14.2	-0.8
DTD	24.3	36.7	-9.4	-1.5
EuroSAT	70.2	9.5	-13.1	-5.6
SUN397	6.8	256.8	-15.5	-2.1
UCF101	21.3	77.4	-14.8	-2.0
ImageNet	5.7	676.5	-18.9	-2.0
ImageNet-A	30.8	178.2	-17.3	-2.7
ImageNet-R	14.5	147.4	-16.2	-2.0
ImageNet-Sketch	6.3	635.7	-14.9	-1.8
ImageNetV2	7.1	706.5	-18.5	-1.9

Backbone: ViT-B/32 Attack: PGD-100

Dataset	P	Rank	Δ_y	$\Delta_{\neg y}$
Aircraft	44.2	93.1	-25.4	-6.3
Cars	49.2	187.0	-37.4	-6.5
Flower102	44.1	93.7	-25.6	-3.8
Food101	62.7	97.2	-26.5	-3.2
Pets	40.2	33.2	-32.5	-6.4
Caltech101	35.4	82.8	-22.6	-2.6
DTD	48.0	41.6	-14.6	-3.0
EuroSAT	93.6	9.9	-20.2	-7.8
SUN397	33.2	354.0	-22.1	-3.0
UCF101	55.9	93.6	-21.8	-3.1
ImageNet	31.8	902.5	-27.7	-3.2
ImageNet-A	66.4	195.8	-25.8	-3.8
ImageNet-R	47.0	181.9	-24.3	-3.4
ImageNet-Sketch	22.8	821.7	-21.8	-3.3
ImageNetV2	33.6	912.8	-27.2	-3.0

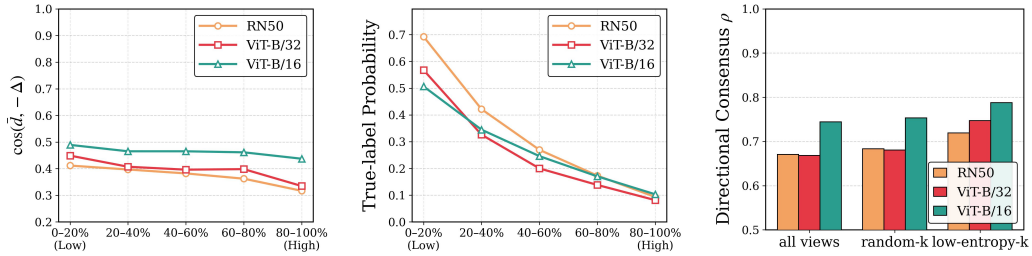
Backbone: ViT-B/32 Attack: PGD-50

Dataset	P	Rank	Δ_y	$\Delta_{\neg y}$
Aircraft	40.1	91.9	-23.6	-5.6
Cars	42.9	184.4	-35.5	-6.2
Flower102	37.6	91.7	-23.9	-3.5
Food101	56.0	95.8	-24.9	-3.0
Pets	33.1	32.2	-29.6	-5.7
Caltech101	28.0	77.3	-20.8	-2.3
DTD	42.3	40.8	-13.5	-2.7
EuroSAT	90.6	9.9	-18.6	-7.3
SUN397	26.1	340.6	-20.7	-2.8
UCF101	48.5	91.6	-20.3	-2.9
ImageNet	24.7	872.6	-25.8	-3.0
ImageNet-A	59.6	194.0	-23.9	-3.6
ImageNet-R	39.6	177.1	-22.5	-3.2
ImageNet-Sketch	18.7	792.6	-20.3	-3.0
ImageNetV2	26.1	884.9	-25.2	-2.9

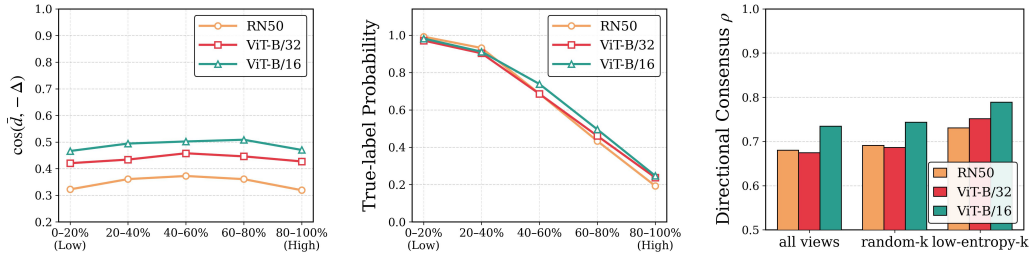
B.2 Detailed Results for RQ2: Efficacy of Entropy Filtering

B.2.1 Entropy Filtering Mechanism Analysis

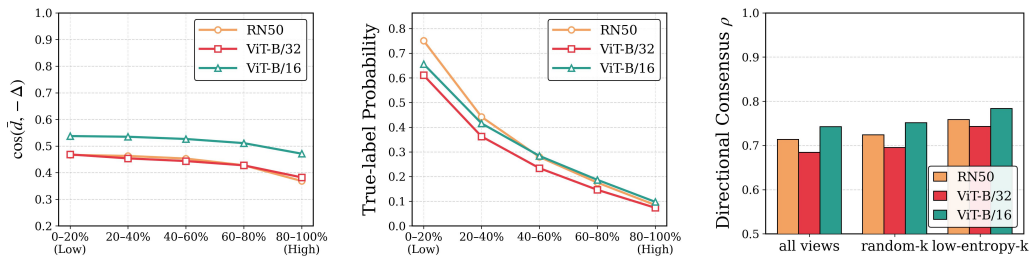
Figures 7–11 are ordered by attack (FGSM, CW, AutoAttack, PGD-10, PGD-100). Within each figure, the first two panels show $\cos(\bar{\mathbf{d}}, -\boldsymbol{\Delta})$ and true-label probability, binned by entropy percentile; low-entropy views consistently yield higher alignment with the reverse attack direction and higher true-label probability. The third panel reports directional consensus ρ under three view-selection strategies: all views, random-k, and low-entropy-k. The low-entropy strategy raises ρ most clearly under PGD-100, where the attack signal is strongest.



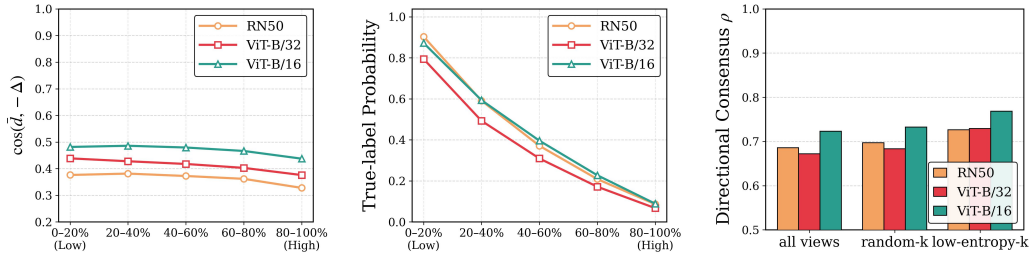
(a) Flower102



(b) Caltech101

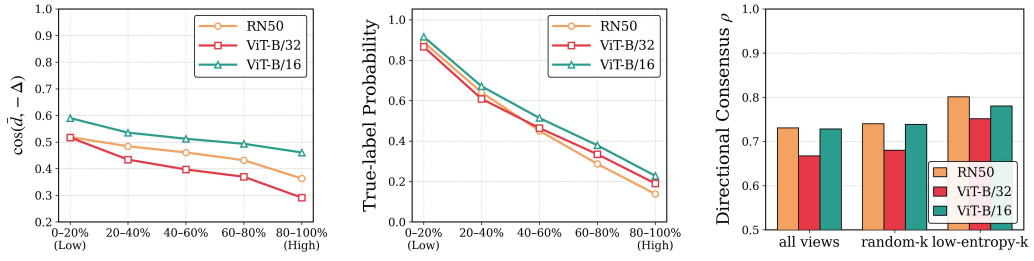


(c) ImageNet

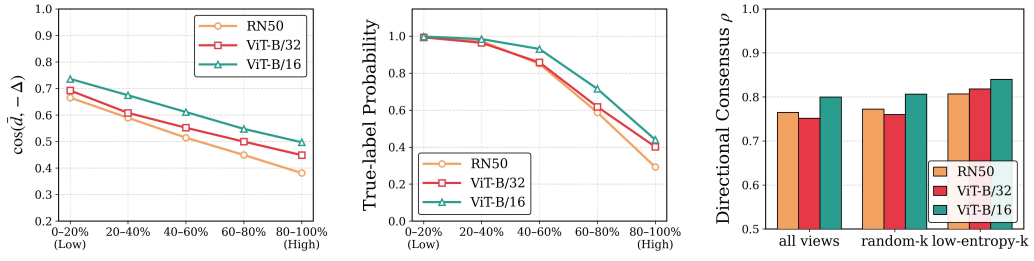


(d) ImageNet-R

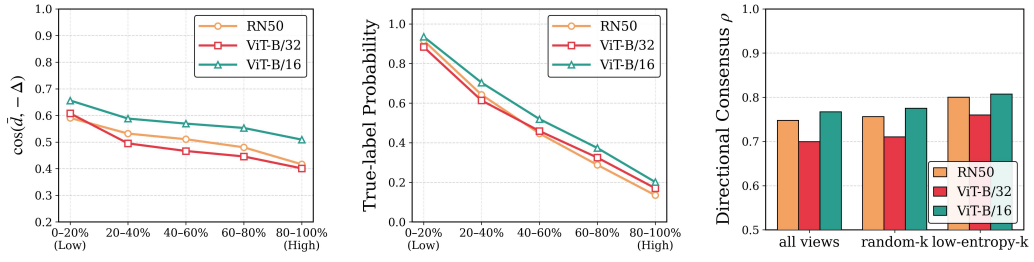
Figure 7: FGSM-attack results for RQ2 on the four representative datasets.



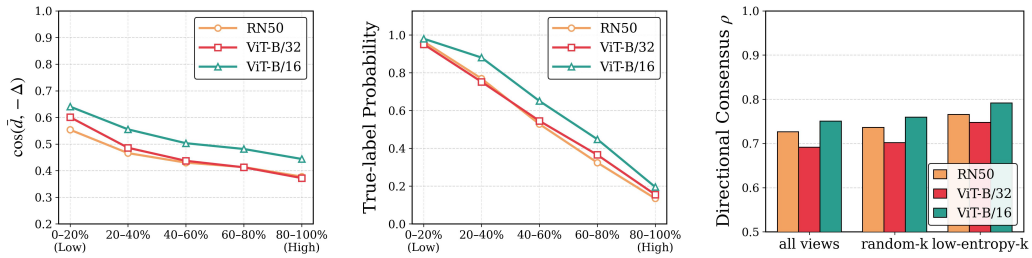
(a) Flower102



(b) Caltech101

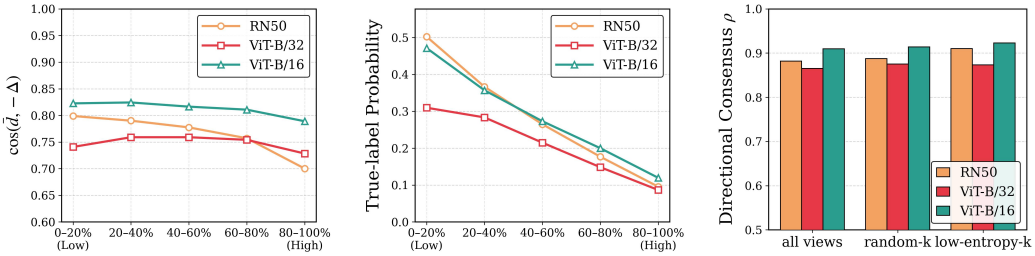


(c) ImageNet

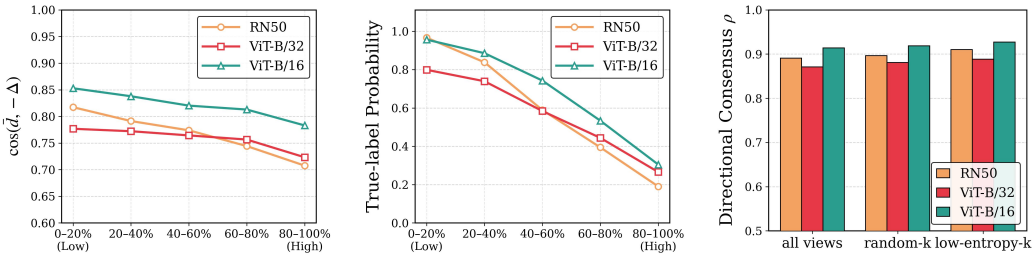


(d) ImageNet-R

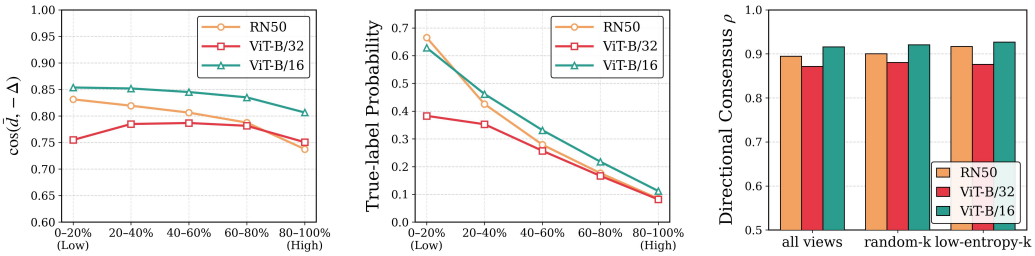
Figure 8: CW-attack results for RQ2 on the four representative datasets.



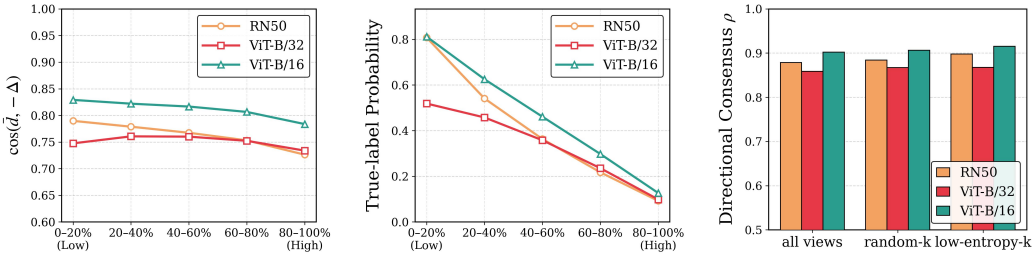
(a) Flower102



(b) Caltech101

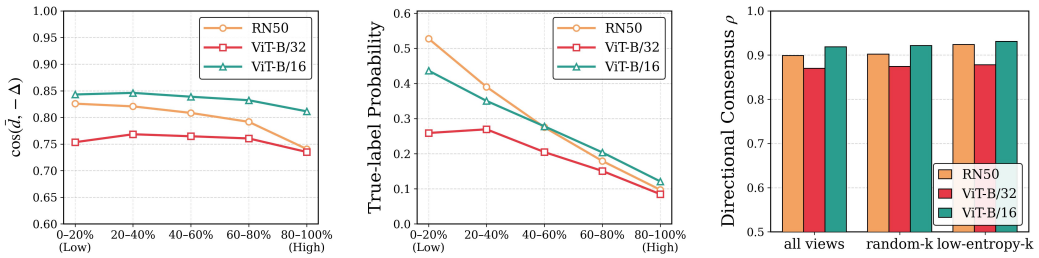


(c) ImageNet

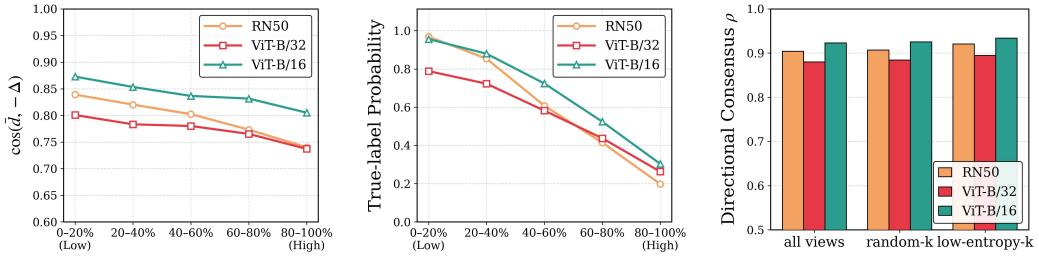


(d) ImageNet-R

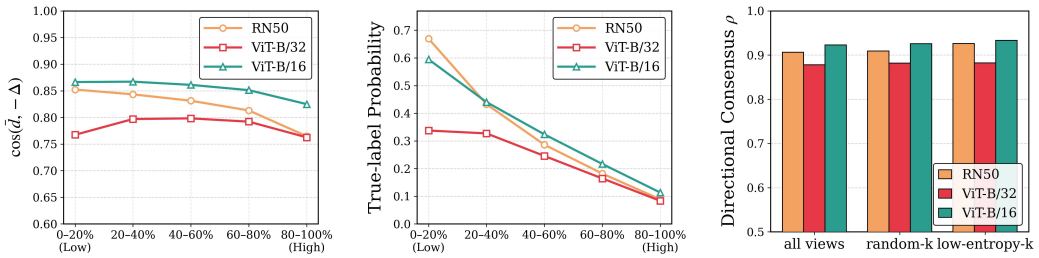
Figure 9: AA-attack results for RQ2 on the four representative datasets.



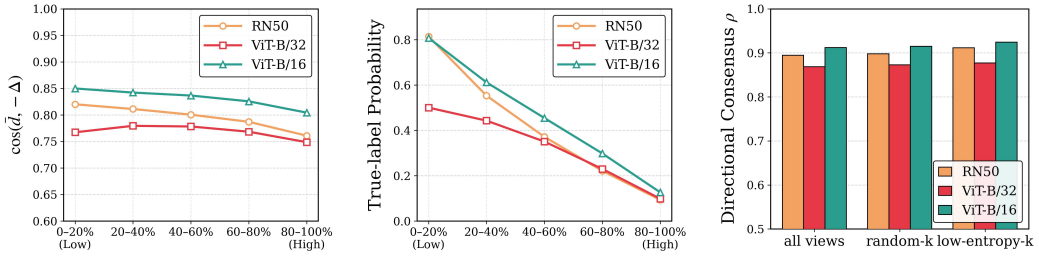
(a) Flower102



(b) Caltech101

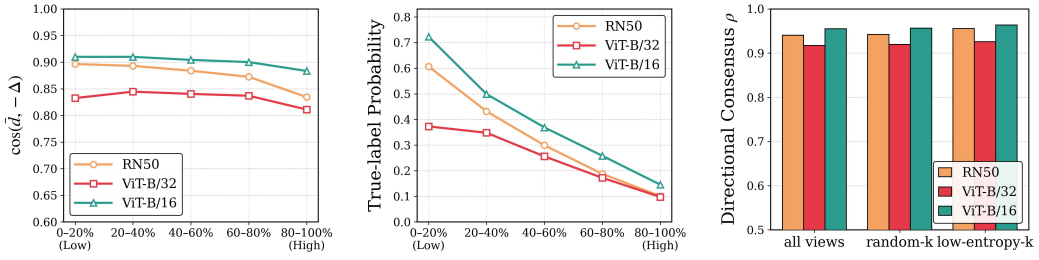


(c) ImageNet

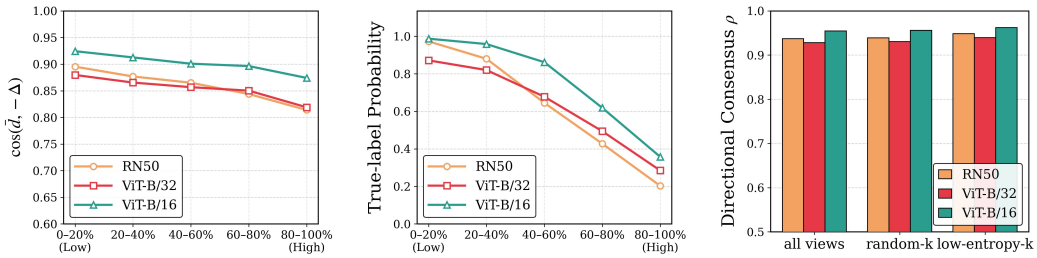


(d) ImageNet-R

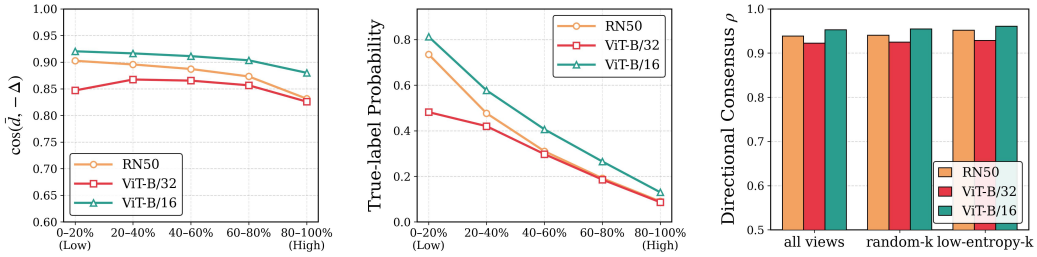
Figure 10: PGD-10-attack results for RQ2 on the four representative datasets.



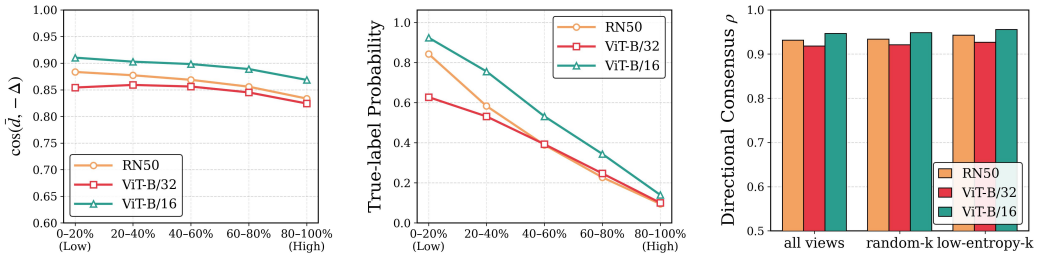
(a) Flower102



(b) Caltech101



(c) ImageNet



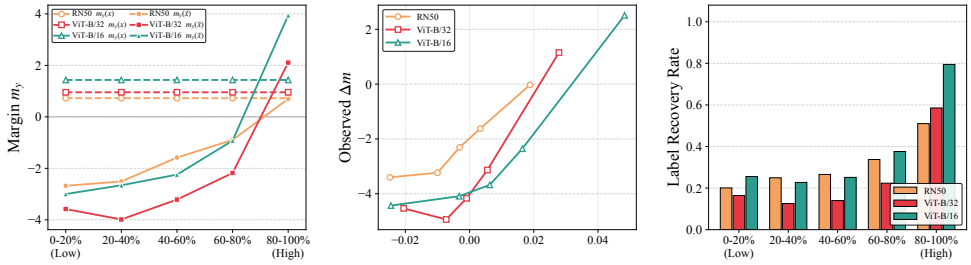
(d) ImageNet-R

Figure 11: PGD-100-attack results for RQ2 on the four representative datasets.

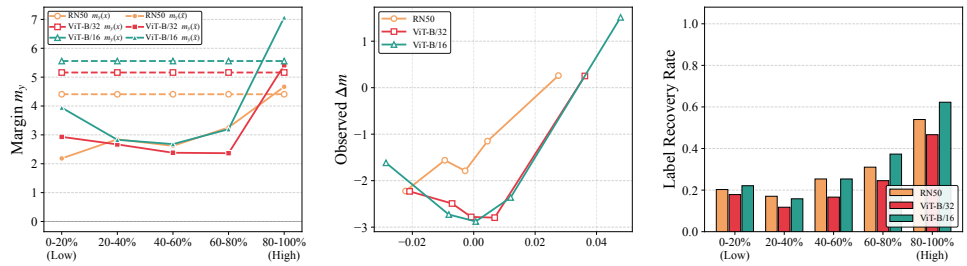
B.3 Detailed Results for RQ3: Directional Extrapolation and Margin Gain

B.3.1 Analysis of Margin Gain

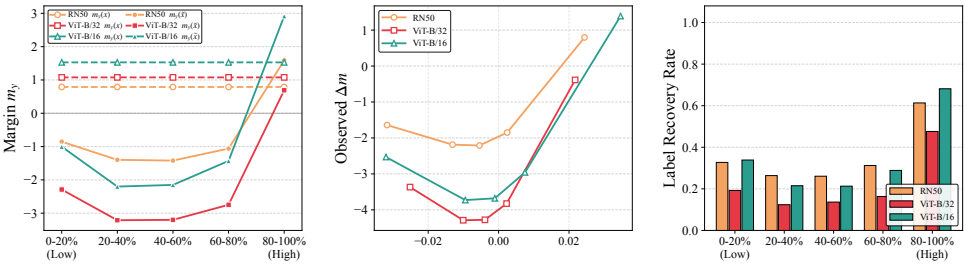
Figures 12–16 follow the same attack order. The left panel compares the clean true-class margin $m_y(\mathbf{x})$ against the DBD-reconstructed margin $m_y(\tilde{\mathbf{x}})$ across predicted-gain bins: DBD lifts the margin beyond the clean baseline, often by a substantial amount. The middle panel relates the theoretical gain G to the observed margin change $\Delta m = m_y(\tilde{\mathbf{x}}) - m_y(\mathbf{x})$, and the right panel stratifies label recovery by gain bin. Stronger attacks with higher G yield higher recovery rates.



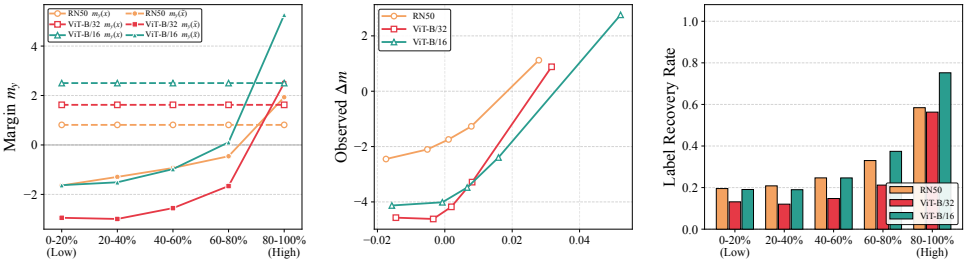
(a) Flower102



(b) Caltech101

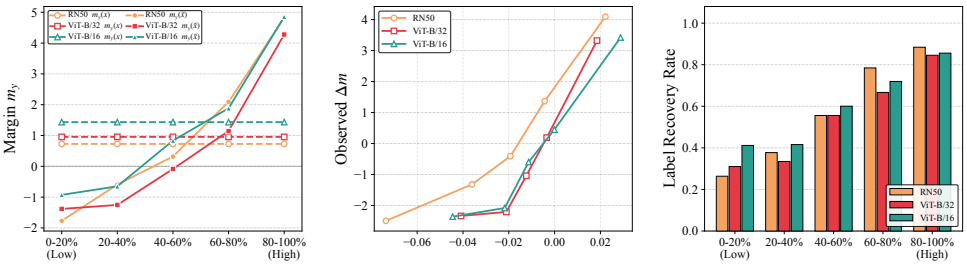


(c) ImageNet

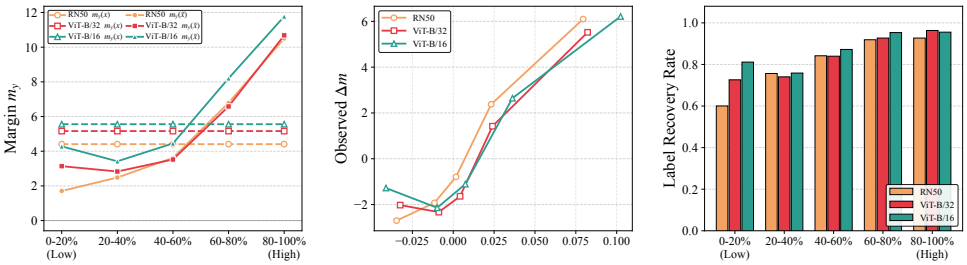


(d) ImageNet-R

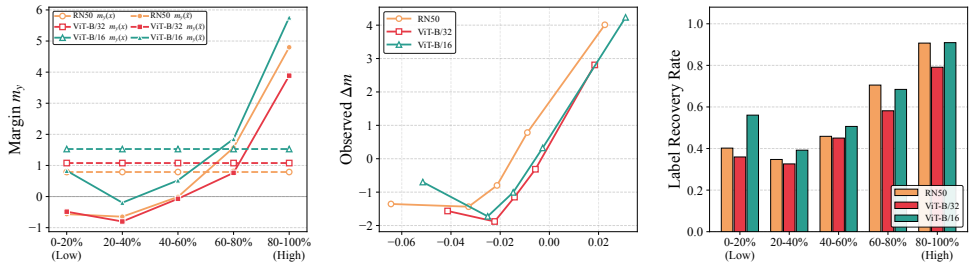
Figure 12: FGSM-attack results for RQ3 on the four representative datasets.



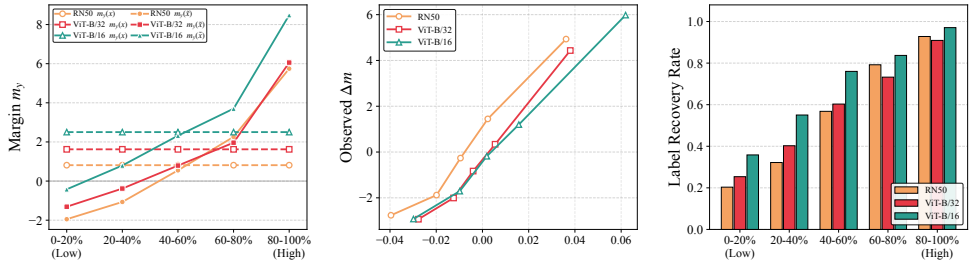
(a) Flower102



(b) Caltech101

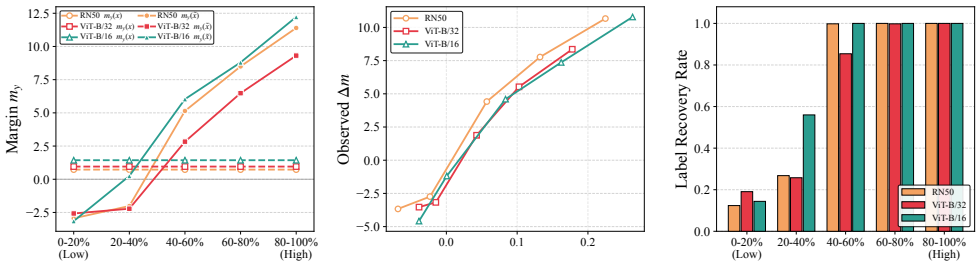


(c) ImageNet

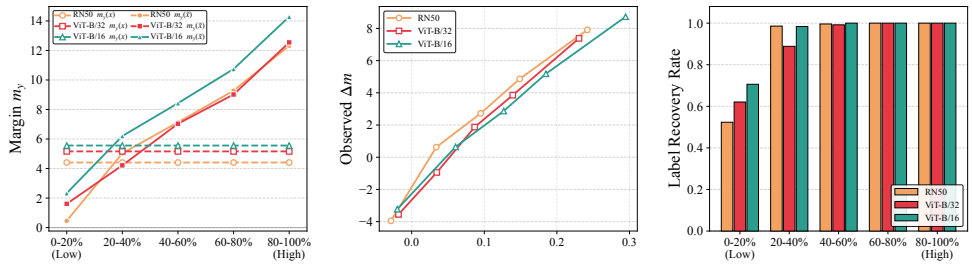


(d) ImageNet-R

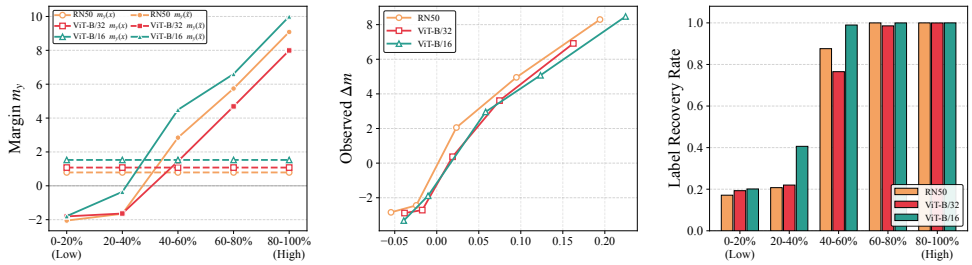
Figure 13: CW-attack results for RQ3 on the four representative datasets.



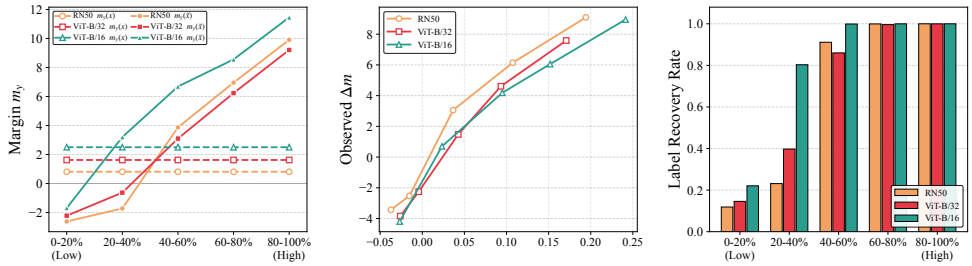
(a) Flower102



(b) Caltech101

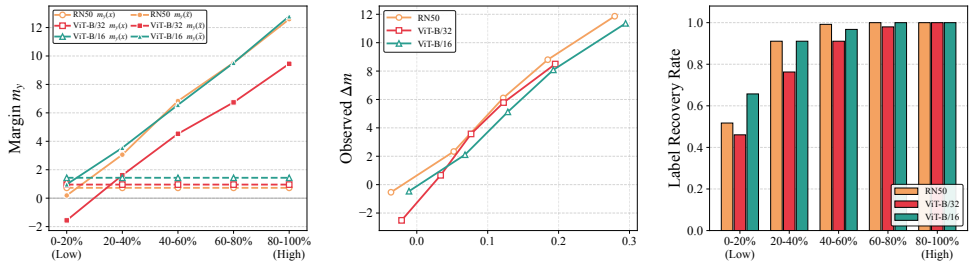


(c) ImageNet

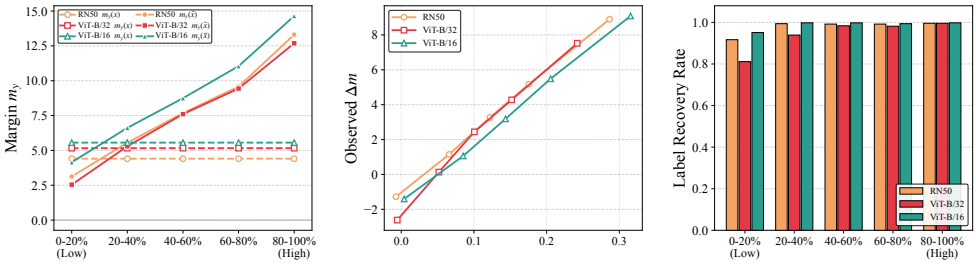


(d) ImageNet-R

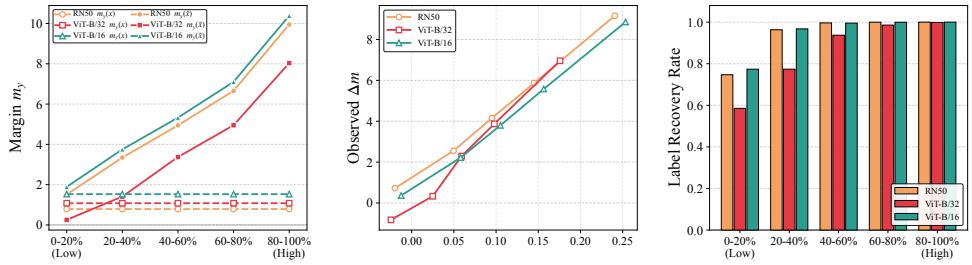
Figure 14: AA-attack results for RQ3 on the four representative datasets.



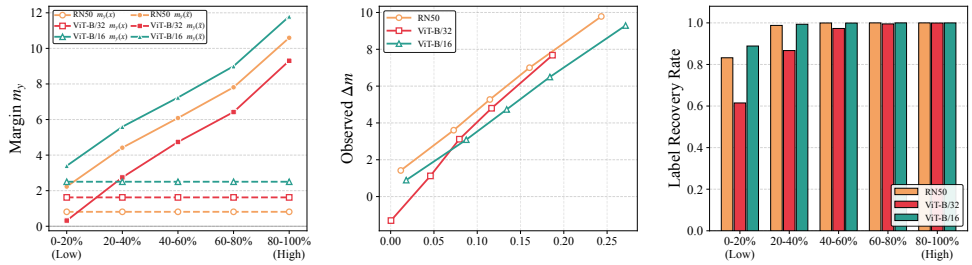
(a) Flower102



(b) Caltech101

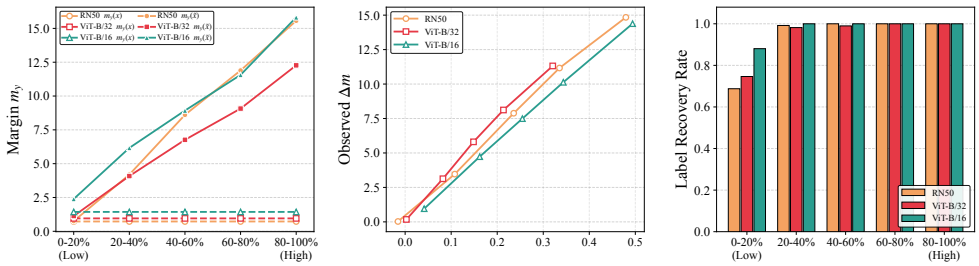


(c) ImageNet

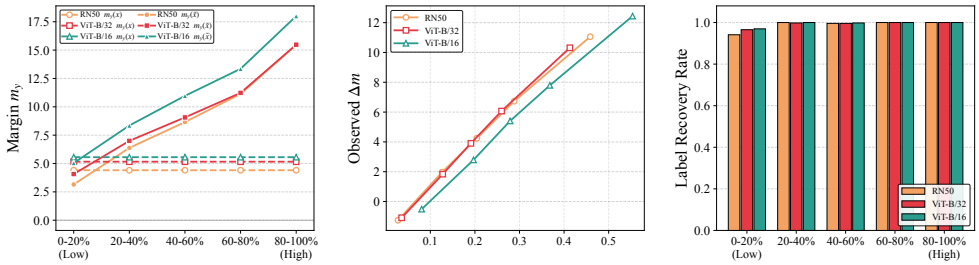


(d) ImageNet-R

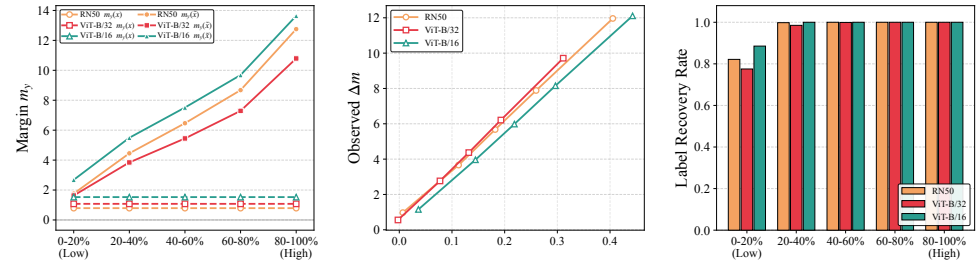
Figure 15: PGD-10-attack results for RQ3 on the four representative datasets.



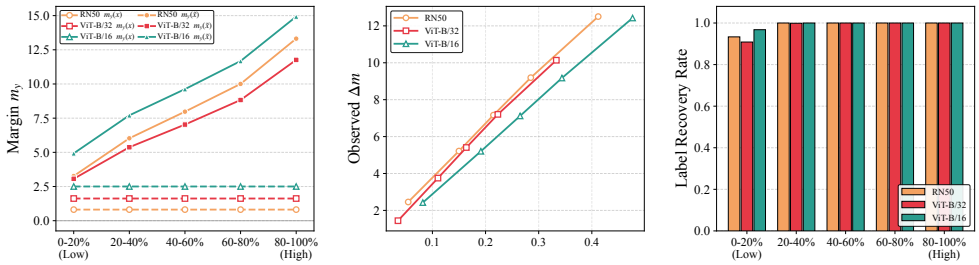
(a) Flower102



(b) Caltech101



(c) ImageNet



(d) ImageNet-R

Figure 16: PGD-100-attack results for RQ3 on the four representative datasets.

B.3.2 Quantitative Metrics of the Margin Gain Mechanism

Table 7 reports dataset-level margin statistics. G is the predicted gain from the DBD update $\beta(\rho)\bar{\mathbf{d}}$; $\Delta m = m_y(\tilde{\mathbf{x}}) - m_y(\mathbf{x})$ is the observed margin lift over the clean margin; LR denotes the label recovery rate (fraction of attacked samples restored to the correct class); Robust is the final defended accuracy. On Flower102 with ViT-B/16 under PGD-100, the mean margin lifts from 1.4 to 9.0, and both LR and Robust reach 97.6%.

Table 7: Dataset-level RQ3 margin-gain mechanism results from all available raw data (all-sample scope).

Backbone: RN50 Attack: FGSM

Dataset	G	Δm	LR	Robust
Aircraft	0.0	-0.6	11.2	11.7
Cars	0.0	-1.6	27.2	28.8
Flower102	0.0	-2.1	33.3	37.3
Food101	0.0	-3.2	31.8	35.7
Pets	0.0	-1.4	56.4	60.9
Caltech101	0.0	-1.3	54.1	74.8
DTD	0.0	-0.6	26.6	38.5
EuroSAT	0.0	-1.0	7.1	17.3
SUN397	0.0	-1.2	37.7	43.9
UCF101	0.0	-1.7	25.0	35.2
ImageNet	0.0	-1.4	38.2	42.6
ImageNet-A	0.0	-2.7	7.5	8.0
ImageNet-R	0.0	-1.3	36.0	44.3
ImageNet-Sketch	0.0	-0.7	21.1	28.1
ImageNetV2	0.0	-1.7	30.2	34.4

Backbone: RN50 Attack: CW

Dataset	G	Δm	LR	Robust
Aircraft	0.0	0.1	17.6	17.7
Cars	0.0	-0.3	49.1	49.3
Flower102	0.0	0.3	57.9	57.8
Food101	0.0	0.0	65.6	65.5
Pets	0.0	0.3	80.1	80.0
Caltech101	0.0	0.6	85.0	85.5
DTD	0.0	0.2	44.8	45.3
EuroSAT	0.0	-0.3	24.1	25.4
SUN397	0.0	0.2	58.5	58.6
UCF101	0.0	0.0	50.7	51.1
ImageNet	0.0	0.2	57.0	57.1
ImageNet-A	0.0	-0.6	16.3	16.3
ImageNet-R	0.0	0.3	57.4	57.9
ImageNet-Sketch	0.0	0.0	33.1	34.4
ImageNetV2	0.0	0.0	49.0	49.1

Backbone: RN50 Attack: AutoAttack

Dataset	G	Δm	LR	Robust
Aircraft	0.0	0.5	22.0	22.0
Cars	0.0	1.8	60.7	60.7
Flower102	0.1	3.3	67.8	67.8
Food101	0.1	2.9	77.9	77.9
Pets	0.1	3.4	85.9	85.9
Caltech101	0.1	2.4	90.2	90.3
DTD	0.0	1.1	50.5	50.5
EuroSAT	0.0	0.4	42.0	42.0
SUN397	0.0	1.6	66.3	66.3
UCF101	0.1	1.5	62.4	62.4
ImageNet	0.0	2.0	65.1	65.1
ImageNet-A	0.0	0.4	26.4	26.4
ImageNet-R	0.1	2.5	65.2	65.2
ImageNet-Sketch	0.0	1.0	39.6	39.6
ImageNetV2	0.0	1.8	58.4	58.4

Backbone: RN50 Attack: PGD-10

Dataset	G	Δm	LR	Robust
Aircraft	0.0	2.0	60.4	60.4
Cars	0.1	3.4	87.8	87.8
Flower102	0.1	5.7	88.4	88.4
Food101	0.2	5.1	96.6	96.6
Pets	0.1	4.8	95.7	95.7
Caltech101	0.1	3.4	98.2	98.2
DTD	0.1	2.9	85.0	85.0
EuroSAT	0.0	1.8	76.0	76.0
SUN397	0.1	3.4	92.9	92.9
UCF101	0.1	3.2	87.5	87.5
ImageNet	0.1	4.5	94.1	94.1
ImageNet-A	0.1	5.9	88.6	88.6
ImageNet-R	0.1	5.4	96.4	96.4
ImageNet-Sketch	0.1	4.2	84.9	84.9
ImageNetV2	0.1	4.7	92.4	92.4

Backbone: RN50 Attack: PGD-100

Dataset	G	Δm	LR	Robust
Aircraft	0.1	2.7	69.7	69.7
Cars	0.1	4.3	91.7	91.7
Flower102	0.2	7.5	93.6	93.6
Food101	0.3	7.3	98.9	98.9
Pets	0.2	6.2	93.7	93.7
Caltech101	0.2	4.6	98.8	98.8
DTD	0.1	3.6	87.9	87.9
EuroSAT	0.1	2.9	84.3	84.3
SUN397	0.2	4.5	96.0	96.0
UCF101	0.2	4.4	92.3	92.3
ImageNet	0.2	6.0	96.4	96.4
ImageNet-A	0.2	8.7	97.7	97.7
ImageNet-R	0.2	7.3	98.7	98.7
ImageNet-Sketch	0.1	5.4	90.6	90.6
ImageNetV2	0.2	6.3	96.0	96.0

Backbone: ViT-B/16 Attack: FGSM

Dataset	G	Δm	LR	Robust
Aircraft	0.0	-1.4	16.3	16.4
Cars	0.0	-2.9	26.3	26.8
Flower102	0.0	-2.4	38.8	39.9
Food101	0.0	-4.3	34.5	37.3
Pets	0.0	-2.6	47.0	49.7
Caltech101	0.0	-1.6	57.7	76.1
DTD	0.0	-1.0	28.6	37.2
EuroSAT	0.0	-2.3	5.5	15.0
SUN397	0.0	-1.7	38.4	42.8
UCF101	0.0	-2.7	26.7	32.8
ImageNet	0.0	-2.3	36.9	40.6
ImageNet-A	0.0	-4.7	12.9	13.8
ImageNet-R	0.0	-2.3	40.4	48.3
ImageNet-Sketch	0.0	-1.6	27.5	33.1
ImageNetV2	0.0	-2.7	30.2	33.5

Backbone: ViT-B/16 Attack: CW

Dataset	G	Δm	LR	Robust
Aircraft	0.0	0.0	23.2	23.3
Cars	0.0	-0.6	49.0	49.1
Flower102	0.0	-0.2	60.2	60.3
Food101	0.0	-0.1	70.7	70.7
Pets	0.0	0.3	80.7	80.7
Caltech101	0.0	0.9	89.8	90.1
DTD	0.0	0.2	47.5	48.0
EuroSAT	0.0	-1.1	25.8	26.7
SUN397	0.0	0.2	60.2	60.3
UCF101	0.0	0.1	56.2	56.3
ImageNet	0.0	0.2	61.4	61.6
ImageNet-A	0.0	-0.6	37.0	37.1
ImageNet-R	0.0	0.5	70.2	70.4
ImageNet-Sketch	0.0	0.0	44.4	45.0
ImageNetV2	0.0	0.1	54.0	54.2

Backbone: ViT-B/16 Attack: AutoAttack

Dataset	G	Δm	LR	Robust
Aircraft	0.0	0.7	30.0	30.0
Cars	0.1	2.4	67.1	67.1
Flower102	0.1	3.4	74.1	74.1
Food101	0.1	2.7	84.8	84.8
Pets	0.1	3.6	89.6	89.6
Caltech101	0.1	2.8	93.9	94.0
DTD	0.0	1.3	55.3	55.3
EuroSAT	0.0	0.2	47.8	47.8
SUN397	0.1	1.8	70.4	70.4
UCF101	0.1	1.8	68.8	68.8
ImageNet	0.1	2.3	72.0	71.9
ImageNet-A	0.0	1.2	53.5	53.5
ImageNet-R	0.1	3.1	80.5	80.5
ImageNet-Sketch	0.0	1.4	52.1	52.1
ImageNetV2	0.1	2.0	65.8	65.8

Backbone: ViT-B/16 Attack: PGD-10

Dataset	G	Δm	LR	Robust
Aircraft	0.0	2.1	65.1	65.1
Cars	0.1	3.8	90.2	90.2
Flower102	0.1	5.2	90.7	90.7
Food101	0.2	4.3	96.7	96.7
Pets	0.1	4.5	96.3	96.3
Caltech101	0.2	3.5	99.0	99.0
DTD	0.1	2.8	85.2	85.2
EuroSAT	0.1	1.3	71.8	71.8
SUN397	0.1	3.3	93.8	93.8
UCF101	0.1	3.2	90.5	90.5
ImageNet	0.1	4.2	94.7	94.7
ImageNet-A	0.1	4.6	91.0	91.0
ImageNet-R	0.1	4.9	97.7	97.7
ImageNet-Sketch	0.1	4.2	90.1	90.1
ImageNetV2	0.1	4.2	93.1	93.1

Backbone: ViT-B/16 Attack: PGD-100

Dataset	G	Δm	LR	Robust
Aircraft	0.1	2.9	72.2	72.2
Cars	0.2	4.9	93.2	93.2
Flower102	0.3	7.5	97.6	97.6
Food101	0.3	7.0	99.5	99.5
Pets	0.3	6.8	96.9	96.9
Caltech101	0.3	5.6	99.4	99.4
DTD	0.1	4.4	91.7	91.7
EuroSAT	0.1	3.7	93.7	93.7
SUN397	0.2	4.9	97.4	97.4
UCF101	0.2	5.1	96.4	96.4
ImageNet	0.2	6.3	97.7	97.7
ImageNet-A	0.3	8.4	98.9	98.9
ImageNet-R	0.3	7.3	99.4	99.4
ImageNet-Sketch	0.1	5.7	95.2	95.2
ImageNetV2	0.2	6.5	97.3	97.3

Backbone: ViT-B/32 Attack: FGSM

Dataset	G	Δm	LR	Robust
Aircraft	0.0	-1.9	8.3	8.4
Cars	0.0	-3.6	14.3	15.0
Flower102	0.0	-3.1	25.7	28.4
Food101	0.0	-5.0	21.4	25.2
Pets	0.0	-3.5	33.0	37.8
Caltech101	0.0	-2.0	44.5	70.8
DTD	0.0	-1.5	19.7	29.8
EuroSAT	0.0	-1.8	9.7	19.6
SUN397	0.0	-2.5	25.9	32.9
UCF101	0.0	-2.6	23.8	31.1
ImageNet	0.0	-3.2	23.4	28.6
ImageNet-A	0.0	-4.9	4.9	5.5
ImageNet-R	0.0	-3.2	26.7	35.5
ImageNet-Sketch	0.0	-2.2	17.6	23.4
ImageNetV2	0.0	-3.5	18.5	23.2

Backbone: ViT-B/32 Attack: CW

Dataset	G	Δm	LR	Robust
Aircraft	0.0	-0.1	16.5	16.6
Cars	0.0	-0.6	43.6	43.7
Flower102	0.0	-0.4	54.7	55.0
Food101	0.0	-0.9	58.3	58.4
Pets	0.0	-0.6	70.5	70.6
Caltech101	0.0	0.2	86.3	86.7
DTD	0.0	-0.1	40.0	40.5
EuroSAT	0.0	-0.8	26.3	27.1
SUN397	0.0	-0.3	52.6	52.8
UCF101	0.0	-0.1	50.4	50.8
ImageNet	0.0	-0.4	50.6	50.8
ImageNet-A	0.0	-0.8	18.9	19.0
ImageNet-R	0.0	-0.2	58.8	59.2
ImageNet-Sketch	0.0	-0.1	37.3	38.2
ImageNetV2	0.0	-0.5	43.2	43.5

Backbone: ViT-B/32 Attack: AutoAttack

Dataset	G	Δm	LR	Robust
Aircraft	0.0	0.3	22.2	22.2
Cars	0.0	1.3	62.0	62.0
Flower102	0.1	1.8	66.0	66.0
Food101	0.1	1.2	75.6	75.6
Pets	0.1	2.2	83.1	83.1
Caltech101	0.1	1.7	90.4	90.4
DTD	0.0	0.6	49.1	49.1
EuroSAT	0.0	-0.2	39.5	39.5
SUN397	0.0	0.9	64.8	64.8
UCF101	0.0	0.9	61.9	61.9
ImageNet	0.0	1.1	63.3	63.3
ImageNet-A	0.0	0.0	29.7	29.7
ImageNet-R	0.1	1.5	68.0	68.0
ImageNet-Sketch	0.0	0.7	42.9	42.9
ImageNetV2	0.0	0.8	56.4	56.4

Backbone: ViT-B/32 Attack: PGD-10

Dataset	G	Δm	LR	Robust
Aircraft	0.0	1.6	53.4	53.4
Cars	0.1	2.4	84.7	84.7
Flower102	0.1	3.2	82.3	82.3
Food101	0.1	2.5	90.0	90.0
Pets	0.1	3.2	94.5	94.5
Caltech101	0.1	2.4	95.1	95.1
DTD	0.0	1.6	75.3	75.3
EuroSAT	0.0	0.6	56.8	56.8
SUN397	0.1	2.0	85.0	85.0
UCF101	0.1	1.9	81.0	81.0
ImageNet	0.1	2.5	85.6	85.6
ImageNet-A	0.1	2.5	67.7	67.7
ImageNet-R	0.1	3.1	89.1	89.1
ImageNet-Sketch	0.0	2.9	77.3	77.3
ImageNetV2	0.1	2.5	82.3	82.3

Backbone: ViT-B/32 Attack: PGD-100

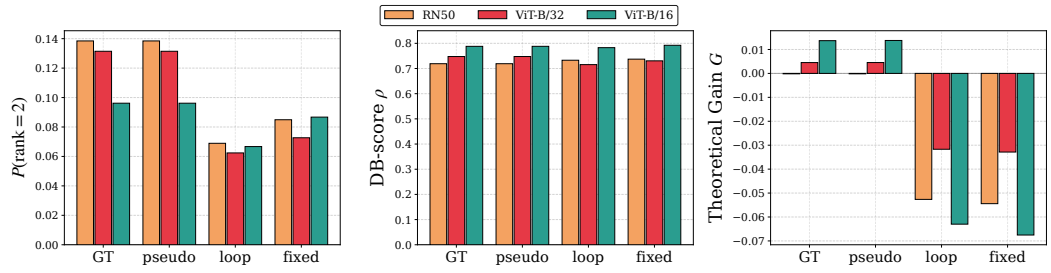
Dataset	G	Δm	LR	Robust
Aircraft	0.0	2.4	66.0	66.0
Cars	0.1	3.6	90.9	90.9
Flower102	0.2	5.7	94.4	94.4
Food101	0.2	5.3	98.4	98.4
Pets	0.2	5.2	96.3	96.3
Caltech101	0.2	4.2	99.3	99.3
DTD	0.1	3.0	87.1	87.1
EuroSAT	0.1	3.0	92.7	92.7
SUN397	0.1	3.5	94.3	94.3
UCF101	0.1	4.0	93.0	93.0
ImageNet	0.1	4.7	95.2	95.2
ImageNet-A	0.2	6.8	95.2	95.2
ImageNet-R	0.2	5.6	98.2	98.2
ImageNet-Sketch	0.1	4.5	89.7	89.7
ImageNetV2	0.1	4.9	94.4	94.4

B.4 Detailed Results for RQ3: Surrogate-Label Ablation

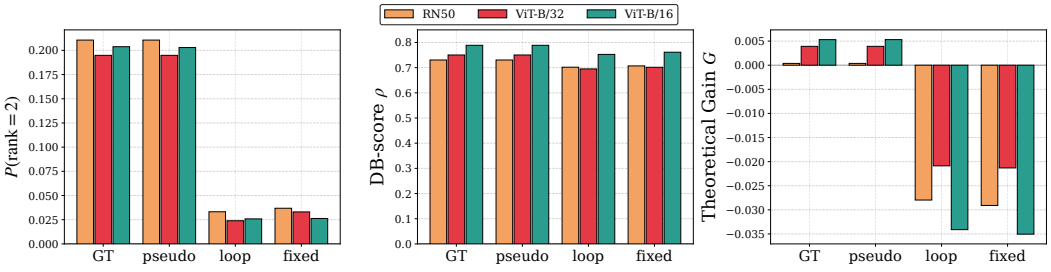
B.4.1 Label Ablation Dynamics on Individual Datasets

Figures 17–21 examine how the three mechanism-level quantities—minimum-logit coverage P, DB-score ρ, and predicted gain G—respond when the attacker's target label is replaced by a surrogate assignment. Flower102 under PGD-100 on ViT-B/16 illustrates the full range: with the ground-truth

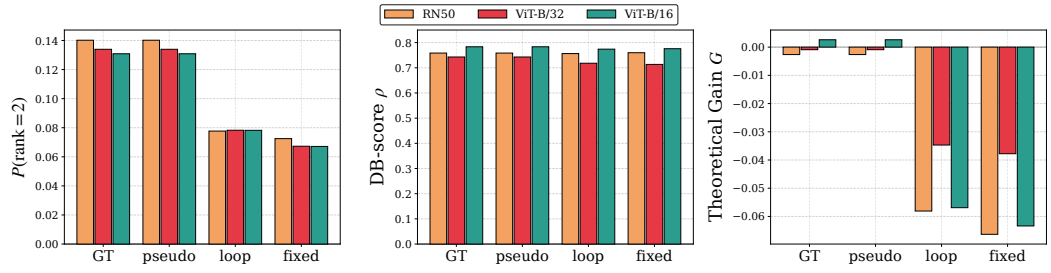
label, DBD Robust is 97.2, $P = 69.7$, and $G = 0.3$; under a fixed-label attack, DBD Robust collapses to 0.2, $P = 0.0$, and $G = -0.7$.



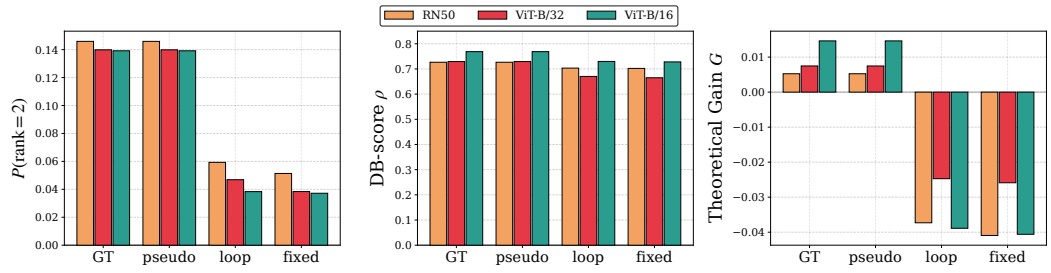
(a) Flower102



(b) Caltech101

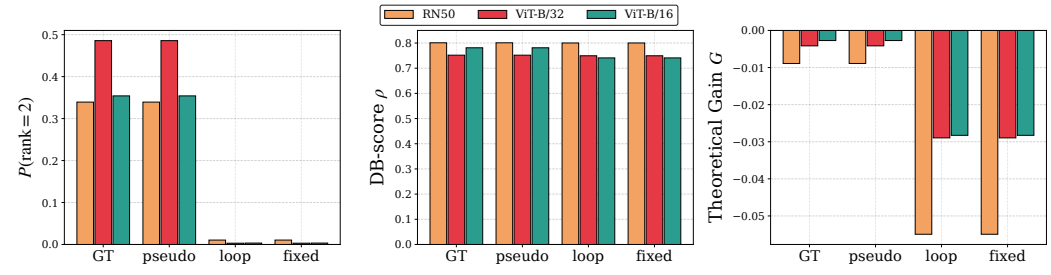


(c) ImageNet

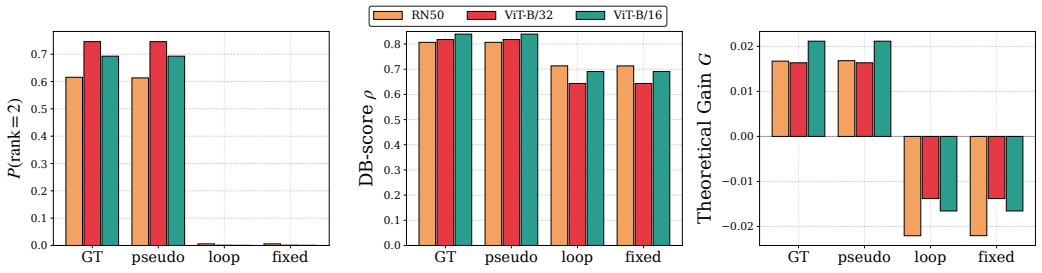


(d) ImageNet-R

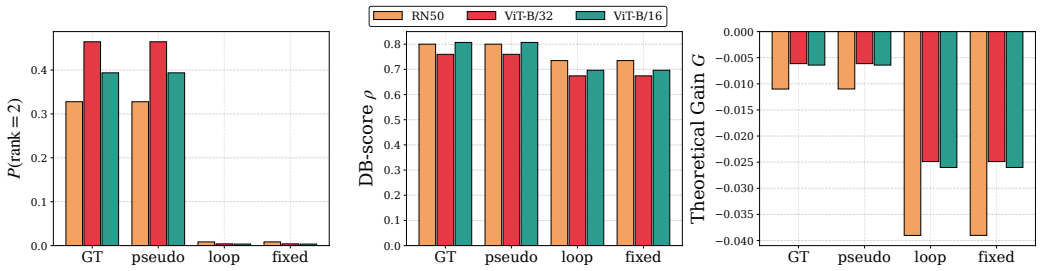
Figure 17: FGSM-attack results for RQ3 on the four representative datasets.



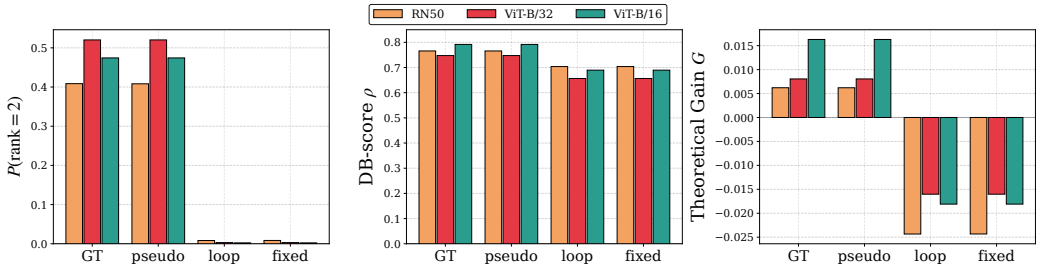
(a) Flower102



(b) Caltech101

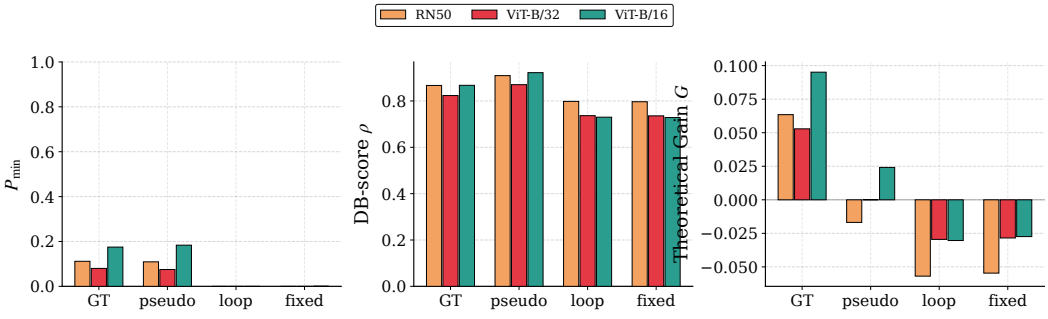


(c) ImageNet

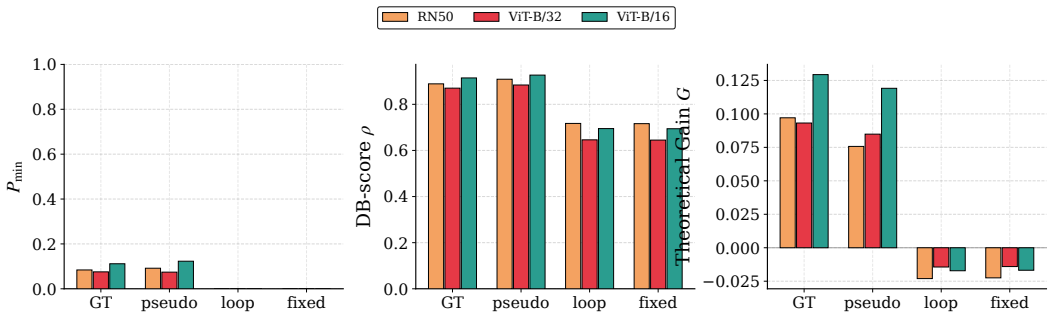


(d) ImageNet-R

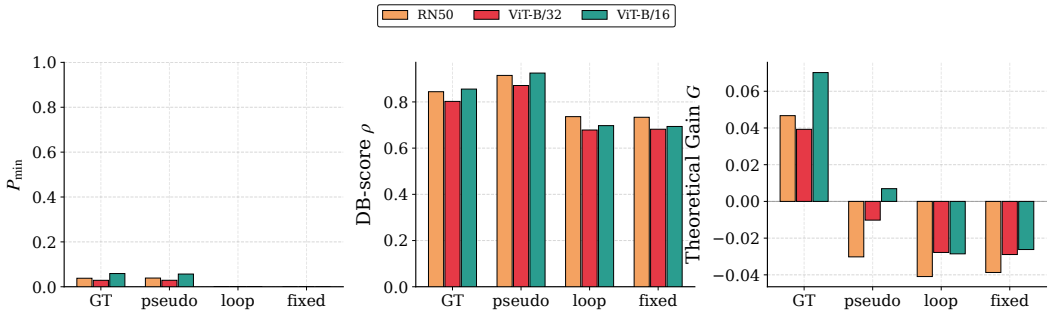
Figure 18: CW-attack results for RQ3 on the four representative datasets.



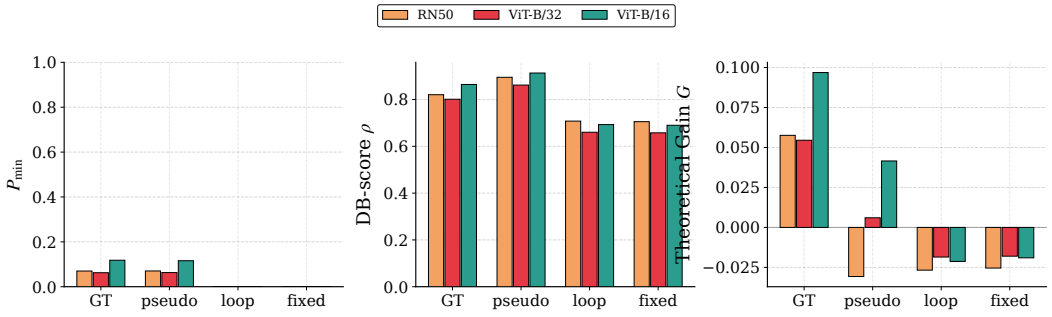
(a) Flower102



(b) Caltech101

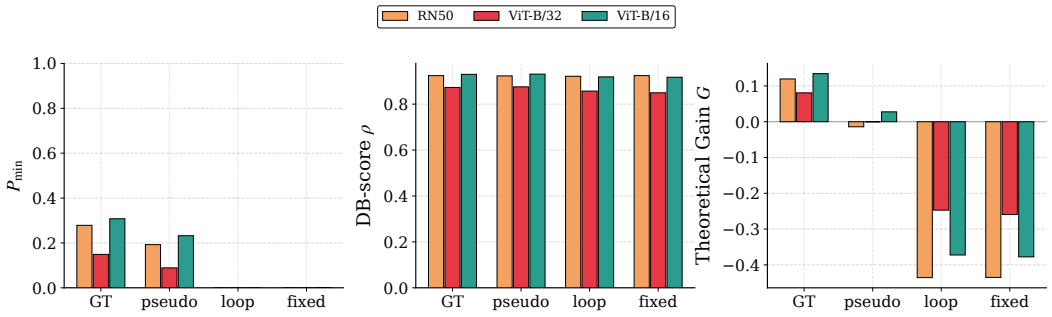


(c) ImageNet

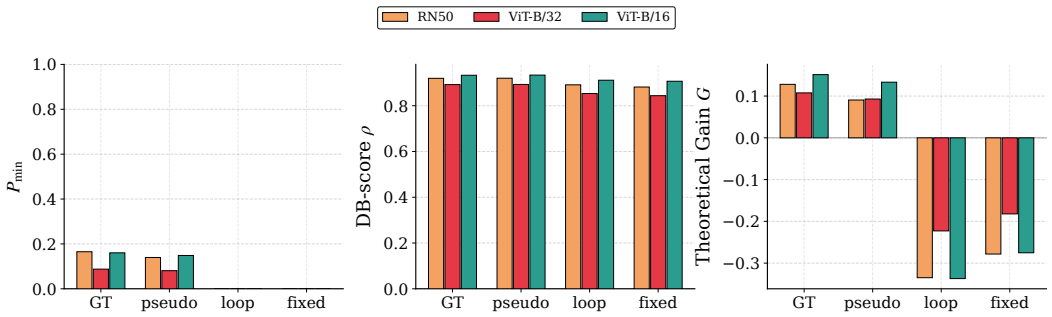


(d) ImageNet-R

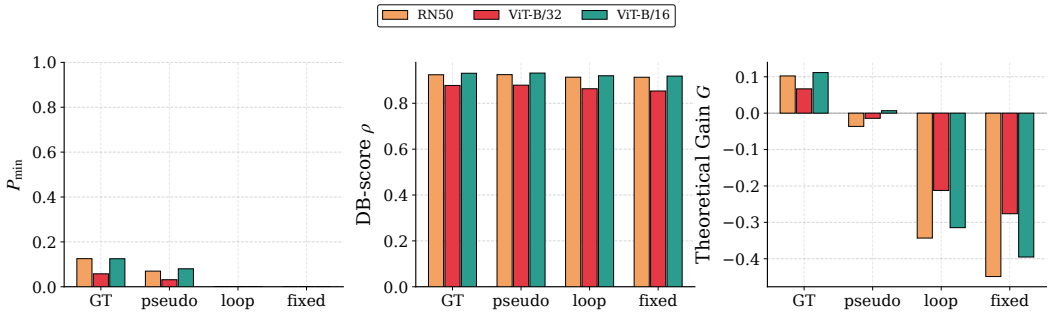
Figure 19: AA-attack results for RQ3 on the four representative datasets.



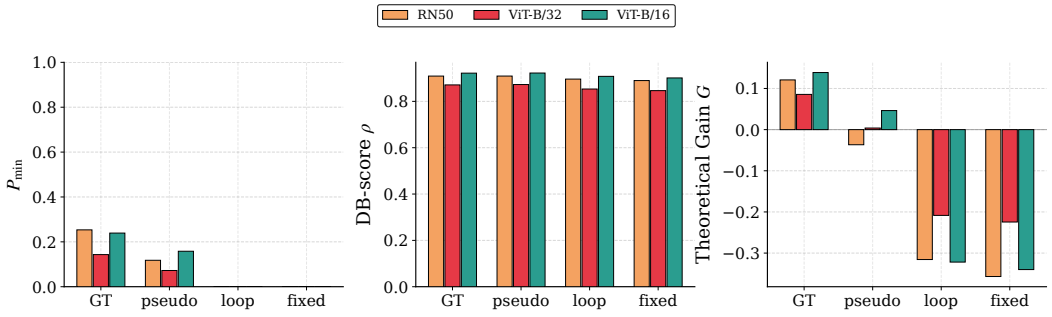
(a) Flower102



(b) Caltech101

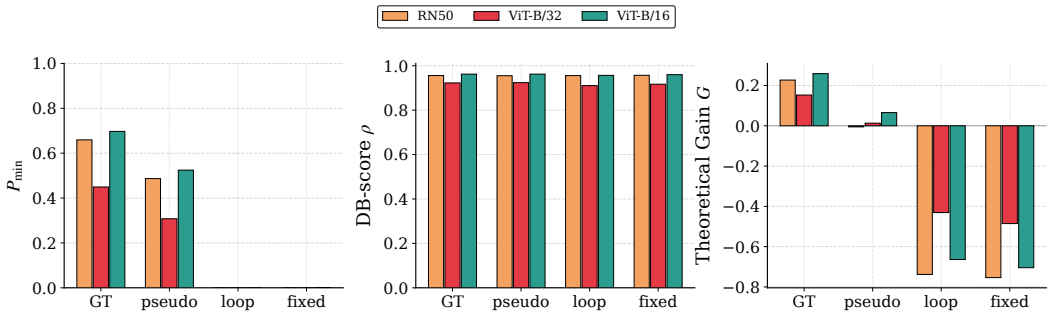


(c) ImageNet

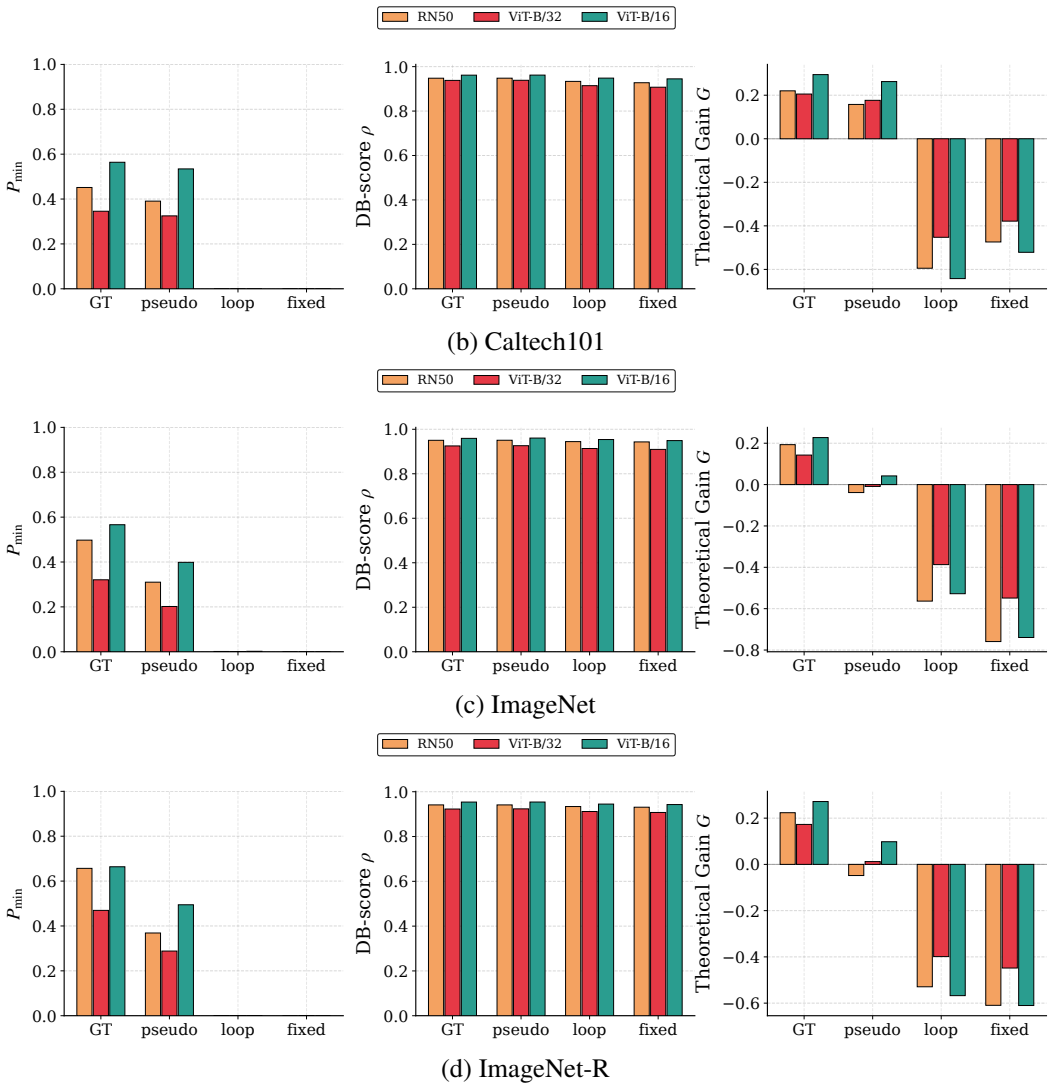


(d) ImageNet-R

Figure 20: PGD-10-attack results for RQ3 on the four representative datasets.



(a) Flower102



(b) Caltech101

(c) ImageNet

(d) ImageNet-R

Figure 21: PGD-100-attack results for RQ3 on the four representative datasets.

B.4.2 Robustness Evaluation with Different Label Assignments

Tables 8–12 split the label-ablation results by attack. Clean denotes the zero-shot CLIP accuracy on the corresponding dataset and backbone, reported as a no-attack reference. DBD is the defended accuracy under adversarial examples constructed with the specified label assignment. The mechanism columns are P, ρ, and G, where P is always computed with respect to the true label. The surrogate-label rows quantify how much of the ground-truth-label behavior persists under other attack labels.

Table 8: Dataset-level RQ3 label-ablation results under the FGSM attack setting from all currently available raw data (all-sample scope). Clean and DBD are reported as percentages.

Backbone: RN50 Label: GT

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	11.8	0.0	0.7	0.0
Cars	57.3	28.7	0.0	0.7	0.0
Flower102	66.0	38.3	0.0	0.7	0.0
Food101	75.5	35.5	0.3	0.8	0.0
Pets	85.0	60.5	0.0	0.7	0.0
Caltech101	89.2	74.8	0.0	0.7	0.0
DTD	48.5	37.8	0.1	0.7	0.0
EuroSAT	37.6	17.5	5.4	0.7	0.0
SUN397	62.8	43.7	0.0	0.8	0.0
UCF101	59.7	34.9	0.0	0.7	0.0
ImageNet	61.5	42.7	0.0	0.8	0.0
ImageNet-A	23.9	7.8	0.3	0.7	0.0
ImageNet-R	60.0	44.4	0.1	0.7	0.0
ImageNet-Sketch	35.6	28.1	0.0	0.7	0.0
ImageNetV2	54.6	34.3	0.0	0.8	0.0

Backbone: ViT-B/16 Label: GT

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	15.9	0.2	0.8	0.0
Cars	66.2	26.6	0.1	0.8	0.0
Flower102	73.0	40.5	0.2	0.8	0.0
Food101	84.2	37.2	0.4	0.8	0.0
Pets	90.3	50.0	0.0	0.8	0.0
Caltech101	94.2	76.1	0.0	0.8	0.0
DTD	53.2	36.8	0.7	0.8	0.0
EuroSAT	55.8	15.0	8.1	0.8	0.0
SUN397	67.9	42.5	0.0	0.8	0.0
UCF101	67.0	33.0	0.1	0.8	0.0
ImageNet	69.7	40.6	0.0	0.8	0.0
ImageNet-A	50.7	13.8	0.5	0.8	0.0
ImageNet-R	77.1	48.4	0.2	0.8	0.0
ImageNet-Sketch	49.1	33.0	0.1	0.7	0.0
ImageNetV2	63.4	33.6	0.0	0.8	0.0

Backbone: RN50 Label: pseudo

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	15.0	0.0	0.7	0.0
Cars	57.3	38.8	0.0	0.7	0.0
Flower102	66.0	44.0	0.0	0.7	0.0
Food101	75.5	40.8	0.0	0.8	0.0
Pets	85.0	65.3	0.0	0.7	0.0
Caltech101	89.2	78.9	0.0	0.7	0.0
DTD	48.5	41.1	0.1	0.7	0.0
EuroSAT	37.6	27.3	3.4	0.7	0.0
SUN397	62.8	49.3	0.0	0.7	0.0
UCF101	59.7	43.0	0.0	0.7	0.0
ImageNet	61.5	46.4	0.0	0.8	0.0
ImageNet-A	23.9	13.2	0.0	0.7	0.0
ImageNet-R	60.0	51.3	0.0	0.7	0.0
ImageNet-Sketch	35.6	32.4	0.0	0.7	0.0
ImageNetV2	54.6	38.8	0.0	0.7	0.0

Backbone: ViT-B/16 Label: pseudo

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	17.3	0.1	0.8	0.0
Cars	66.4	32.2	0.0	0.8	0.0
Flower102	73.0	43.7	0.0	0.8	0.0
Food101	82.8	31.9	0.3	0.8	0.0
Pets	90.3	54.0	0.0	0.8	0.0
Caltech101	94.2	79.2	0.0	0.8	0.0
DTD	53.2	42.3	0.2	0.8	0.0
EuroSAT	44.9	12.1	3.9	0.8	0.0
SUN397	66.5	47.6	0.0	0.8	0.0
UCF101	67.0	41.7	0.0	0.8	0.0
ImageNet	69.7	46.2	0.0	0.8	0.0
ImageNet-A	54.7	25.0	0.2	0.8	0.0
ImageNet-R	77.1	55.6	0.0	0.8	0.0
ImageNet-Sketch	49.1	38.3	0.0	0.8	0.0
ImageNetV2	63.4	39.7	0.0	0.8	0.0

Backbone: RN50 Label: loop

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	14.8	0.0	0.7	0.0
Cars	57.3	51.8	0.0	0.7	-0.1
Flower102	66.0	59.1	0.0	0.7	0.0
Food101	75.5	69.7	0.0	0.8	0.0
Pets	85.0	81.1	0.0	0.8	0.0
Caltech101	89.2	89.3	0.0	0.7	0.0
DTD	48.5	47.2	0.0	0.7	0.0
EuroSAT	37.6	35.3	2.2	0.7	0.0
SUN397	62.8	58.7	0.0	0.7	0.0
UCF101	59.7	57.3	0.0	0.7	0.0
ImageNet	61.5	55.6	0.0	0.8	-0.1
ImageNet-A	23.9	17.7	0.0	0.7	0.0
ImageNet-R	60.0	58.9	0.0	0.7	0.0
ImageNet-Sketch	35.6	34.5	0.0	0.7	0.0
ImageNetV2	54.6	50.1	0.0	0.7	0.0

Backbone: ViT-B/16 Label: loop

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	19.6	0.0	0.7	0.0
Cars	66.1	56.5	0.0	0.8	-0.1
Flower102	73.0	67.3	0.0	0.8	-0.1
Food101	82.9	80.4	0.0	0.8	0.0
Pets	88.2	84.6	0.0	0.8	-0.1
Caltech101	94.2	94.4	0.0	0.8	0.0
DTD	53.2	51.5	0.0	0.8	0.0
EuroSAT	55.8	46.4	1.6	0.8	0.0
SUN397	67.1	64.5	0.0	0.8	0.0
UCF101	67.0	64.5	0.0	0.8	0.0
ImageNet	69.7	65.3	0.0	0.8	-0.1
ImageNet-A	50.7	46.1	0.0	0.8	0.0
ImageNet-R	77.1	76.2	0.0	0.7	0.0
ImageNet-Sketch	49.1	46.9	0.0	0.7	0.0
ImageNetV2	63.4	59.6	0.0	0.8	-0.1

Backbone: RN50 Label: fixed

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	18.5	0.0	0.7	0.0
Cars	57.3	55.3	0.0	0.7	-0.1
Flower102	66.0	59.5	0.0	0.7	0.0
Food101	75.5	70.2	0.0	0.8	0.0
Pets	85.0	81.9	0.0	0.8	-0.1
Caltech101	89.2	89.1	0.0	0.7	0.0
DTD	48.5	47.0	0.0	0.7	0.0
EuroSAT	37.6	35.4	2.8	0.7	0.0
SUN397	62.8	60.0	0.0	0.7	0.0
UCF101	59.7	57.9	0.0	0.7	0.0
ImageNet	61.5	57.2	0.0	0.8	-0.1
ImageNet-A	23.9	20.4	0.0	0.7	0.0
ImageNet-R	60.0	61.1	0.0	0.7	0.0
ImageNet-Sketch	35.6	35.7	0.0	0.7	0.0
ImageNetV2	54.6	51.3	0.0	0.7	-0.1

Backbone: ViT-B/16 Label: fixed

Dataset	Clean	DBD	P	ρ	G
Aircraft	24.6	23.5	0.0	0.7	0.0
Cars	66.2	60.9	0.0	0.7	-0.1
Flower102	73.0	68.0	0.0	0.8	-0.1
Food101	84.2	82.7	0.0	0.8	0.0
Pets	88.7	87.4	0.0	0.8	-0.1
Caltech101	94.2	94.0	0.0	0.8	0.0
DTD	53.2	49.5	0.0	0.8	0.0
EuroSAT	55.8	49.2	1.5	0.8	0.0
SUN397	67.9	65.4	0.0	0.8	0.0
UCF101	64.4	62.5	0.0	0.8	0.0
ImageNet	69.7	67.8	0.0	0.8	-0.1
ImageNet-A	50.7	50.4	0.0	0.8	0.0
ImageNet-R	77.1	78.3	0.0	0.7	0.0
ImageNet-Sketch	49.1	49.1	0.0	0.7	0.0
ImageNetV2	63.4	61.6	0.0	0.8	-0.1

Backbone: ViT-B/32 Label: GT

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	8.5	0.2	0.7	0.0
Cars	61.2	14.8	0.0	0.7	0.0
Flower102	67.0	28.4	0.0	0.7	0.0
Food101	78.7	25.4	0.3	0.7	0.0
Pets	86.6	37.5	0.0	0.7	0.0
Caltech101	93.3	71.2	0.0	0.7	0.0
DTD	49.9	29.7	0.2	0.7	0.0
EuroSAT	50.8	19.4	7.1	0.8	0.0
SUN397	65.7	32.9	0.0	0.7	0.0
UCF101	63.6	31.1	0.0	0.7	0.0
ImageNet	64.4	28.6	0.0	0.7	0.0
ImageNet-A	31.1	5.2	0.5	0.7	0.0
ImageNet-R	68.3	35.7	0.2	0.7	0.0
ImageNet-Sketch	42.5	23.4	0.1	0.7	0.0
ImageNetV2	57.2	23.4	0.0	0.7	0.0

Backbone: ViT-B/32 Label: loop

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	14.1	0.0	0.7	0.0
Cars	61.2	51.8	0.0	0.7	0.0
Flower102	67.0	64.8	0.0	0.7	0.0
Food101	76.6	74.7	0.0	0.7	0.0
Pets	86.6	84.4	0.0	0.7	0.0
Caltech101	93.3	93.4	0.0	0.7	0.0
DTD	49.9	49.0	0.0	0.7	0.0
EuroSAT	50.8	44.7	2.7	0.7	0.0
SUN397	65.7	62.7	0.0	0.7	0.0
UCF101	63.6	62.9	0.0	0.7	0.0
ImageNet	64.4	60.7	0.0	0.7	0.0
ImageNet-A	31.1	27.7	0.1	0.7	0.0
ImageNet-R	68.3	67.3	0.0	0.7	0.0
ImageNet-Sketch	42.5	40.2	0.0	0.7	0.0
ImageNetV2	57.2	54.1	0.0	0.7	0.0

Backbone: ViT-B/32 Label: pseudo

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	13.5	0.0	0.7	0.0
Cars	61.2	24.5	0.0	0.7	0.0
Flower102	67.0	35.8	0.0	0.7	0.0
Food101	78.7	32.8	0.1	0.7	0.0
Pets	86.6	43.9	0.0	0.7	0.0
Caltech101	93.3	75.0	0.0	0.7	0.0
DTD	49.9	36.1	0.1	0.8	0.0
EuroSAT	42.7	31.0	1.4	0.8	0.0
SUN397	65.7	43.0	0.0	0.7	0.0
UCF101	63.6	41.6	0.0	0.7	0.0
ImageNet	64.4	37.7	0.0	0.7	0.0
ImageNet-A	31.1	17.5	0.0	0.7	0.0
ImageNet-R	68.3	45.6	0.0	0.7	0.0
ImageNet-Sketch	42.5	31.3	0.0	0.7	0.0
ImageNetV2	57.2	33.3	0.0	0.7	0.0

Backbone: ViT-B/32 Label: fixed

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	20.4	0.0	0.7	0.0
Cars	61.2	58.8	0.0	0.7	0.0
Flower102	67.0	64.8	0.0	0.7	0.0
Food101	78.7	77.9	0.0	0.7	0.0
Pets	86.6	85.9	0.0	0.7	0.0
Caltech101	93.3	92.7	0.0	0.7	0.0
DTD	49.9	48.1	0.0	0.7	0.0
EuroSAT	50.8	49.2	1.6	0.7	0.0
SUN397	65.7	64.8	0.0	0.7	0.0
UCF101	63.6	63.3	0.0	0.7	0.0
ImageNet	64.4	64.0	0.0	0.7	0.0
ImageNet-A	31.1	31.7	0.0	0.7	0.0
ImageNet-R	68.3	70.5	0.0	0.7	0.0
ImageNet-Sketch	42.5	43.0	0.0	0.7	0.0
ImageNetV2	57.2	56.7	0.0	0.7	0.0

Table 9: Dataset-level RQ3 label-ablation results under the CW attack setting from all currently available raw data (all-sample scope). Clean and DBD are reported as percentages.

Backbone: RN50 Label: GT

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	17.3	0.0	0.7	0.0
Cars	57.3	49.1	0.0	0.7	0.0
Flower102	65.9	57.7	0.0	0.8	0.0
Food101	75.5	65.4	0.0	0.8	0.0
Pets	85.1	80.0	0.0	0.8	0.0
Caltech101	89.2	85.6	0.0	0.8	0.0
DTD	48.4	45.3	0.0	0.8	0.0
EuroSAT	37.6	25.5	2.3	0.7	0.0
SUN397	62.8	58.5	0.0	0.8	0.0
UCF101	59.7	51.4	0.0	0.7	0.0
ImageNet	61.5	57.0	0.0	0.8	0.0
ImageNet-A	23.9	16.3	0.0	0.7	0.0
ImageNet-R	60.0	57.8	0.0	0.7	0.0
ImageNet-Sketch	35.6	34.4	0.0	0.7	0.0
ImageNetV2	54.6	49.4	0.0	0.8	0.0

Backbone: RN50 Label: loop

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	19.5	0.0	0.7	0.0
Cars	57.3	57.9	0.0	0.7	0.0
Flower102	65.9	57.1	0.0	0.8	-0.1
Food101	75.6	67.0	0.0	0.8	0.0
Pets	85.1	84.1	0.0	0.8	0.0
Caltech101	89.2	89.0	0.0	0.7	0.0
DTD	48.5	45.4	0.0	0.8	0.0
EuroSAT	37.5	37.9	1.9	0.6	0.0
SUN397	62.8	61.5	0.0	0.7	0.0
UCF101	59.7	59.3	0.0	0.7	0.0
ImageNet	61.5	60.1	0.0	0.7	0.0
ImageNet-A	23.9	19.4	0.0	0.7	0.0
ImageNet-R	60.0	61.5	0.0	0.7	0.0
ImageNet-Sketch	35.6	36.5	0.0	0.7	0.0
ImageNetV2	54.6	52.1	0.0	0.7	0.0

Backbone: RN50 Label: pseudo

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	18.4	0.0	0.8	0.0
Cars	57.3	52.8	0.0	0.7	0.0
Flower102	65.9	58.3	0.0	0.8	0.0
Food101	75.5	65.8	0.0	0.8	0.0
Pets	85.1	81.2	0.0	0.8	0.0
Caltech101	89.2	86.6	0.0	0.8	0.0
DTD	48.4	46.0	0.0	0.8	0.0
EuroSAT	37.6	31.6	1.7	0.7	0.0
SUN397	62.8	59.4	0.0	0.8	0.0
UCF101	59.7	54.3	0.0	0.8	0.0
ImageNet	61.5	57.4	0.0	0.8	0.0
ImageNet-A	23.9	17.4	0.0	0.8	-0.1
ImageNet-R	60.0	59.2	0.0	0.8	0.0
ImageNet-Sketch	35.6	35.3	0.0	0.7	0.0
ImageNetV2	54.6	50.0	0.0	0.8	0.0

Backbone: RN50 Label: fixed

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	19.3	0.0	0.7	0.0
Cars	57.3	57.5	0.0	0.7	0.0
Flower102	65.9	57.1	0.0	0.8	-0.1
Food101	75.5	67.0	0.0	0.8	0.0
Pets	85.1	84.3	0.0	0.8	0.0
Caltech101	89.2	88.9	0.0	0.7	0.0
DTD	48.4	45.5	0.0	0.8	0.0
EuroSAT	37.6	37.2	2.0	0.6	0.0
SUN397	62.8	61.4	0.0	0.7	0.0
UCF101	59.7	59.1	0.0	0.7	0.0
ImageNet	61.5	60.1	0.0	0.7	0.0
ImageNet-A	23.9	19.5	0.0	0.7	0.0
ImageNet-R	60.0	61.4	0.0	0.7	0.0
ImageNet-Sketch	35.6	36.4	0.0	0.7	0.0
ImageNetV2	54.6	52.0	0.0	0.7	0.0

Backbone: ViT-B/16 Label: GT

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	23.4	0.0	0.7	0.0
Cars	66.2	48.7	0.0	0.7	0.0
Flower102	73.0	59.9	0.0	0.8	0.0
Food101	84.2	70.8	0.0	0.8	0.0
Pets	90.3	80.7	0.0	0.8	0.0
Caltech101	94.2	90.1	0.0	0.8	0.0
DTD	53.2	48.0	0.0	0.8	0.0
EuroSAT	55.8	26.6	2.7	0.8	0.0
SUN397	67.9	60.0	0.0	0.8	0.0
UCF101	67.0	56.3	0.0	0.8	0.0
ImageNet	69.7	61.5	0.0	0.8	0.0
ImageNet-A	50.7	37.1	0.0	0.7	0.0
ImageNet-R	77.1	70.4	0.0	0.8	0.0
ImageNet-Sketch	49.1	45.1	0.0	0.7	0.0
ImageNetV2	63.4	54.1	0.0	0.8	0.0

Backbone: ViT-B/32 Label: GT

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	16.3	0.0	0.7	0.0
Cars	61.2	43.4	0.0	0.7	0.0
Flower102	67.0	55.7	0.0	0.7	0.0
Food101	78.7	58.4	0.0	0.7	0.0
Pets	86.6	71.0	0.0	0.8	0.0
Caltech101	93.3	86.5	0.0	0.8	0.0
DTD	49.9	41.1	0.0	0.7	0.0
EuroSAT	50.8	27.0	1.7	0.7	0.0
SUN397	65.7	52.7	0.0	0.7	0.0
UCF101	63.6	50.8	0.0	0.8	0.0
ImageNet	64.4	50.8	0.0	0.7	0.0
ImageNet-A	31.1	18.8	0.0	0.7	0.0
ImageNet-R	68.3	59.4	0.0	0.7	0.0
ImageNet-Sketch	42.5	38.0	0.0	0.7	0.0
ImageNetV2	57.2	43.6	0.0	0.7	0.0

Backbone: ViT-B/16 Label: pseudo

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	25.0	0.0	0.8	0.0
Cars	66.2	54.2	0.0	0.8	0.0
Flower102	73.0	61.4	0.0	0.8	0.0
Food101	84.2	72.1	0.0	0.8	0.0
Pets	90.3	82.0	0.0	0.8	0.0
Caltech101	94.2	91.3	0.0	0.8	0.0
DTD	53.2	50.2	0.0	0.8	0.0
EuroSAT	55.8	32.2	3.7	0.8	0.0
SUN397	67.9	63.0	0.0	0.8	0.0
UCF101	67.0	59.4	0.0	0.8	0.0
ImageNet	69.7	63.7	0.0	0.8	0.0
ImageNet-A	50.7	40.1	0.0	0.8	0.0
ImageNet-R	77.1	73.0	0.0	0.8	0.0
ImageNet-Sketch	49.1	47.6	0.0	0.7	0.0
ImageNetV2	63.4	57.0	0.0	0.8	0.0

Backbone: ViT-B/32 Label: pseudo

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	19.3	0.0	0.7	0.0
Cars	61.2	49.6	0.0	0.7	0.0
Flower102	67.0	59.9	0.0	0.7	0.0
Food101	78.7	62.9	0.0	0.8	0.0
Pets	86.6	75.0	0.0	0.8	0.0
Caltech101	93.3	88.2	0.0	0.8	0.0
DTD	49.9	44.9	0.0	0.8	0.0
EuroSAT	50.8	36.6	1.6	0.8	0.0
SUN397	65.7	58.4	0.0	0.8	0.0
UCF101	63.6	56.7	0.0	0.8	0.0
ImageNet	64.4	56.1	0.0	0.7	0.0
ImageNet-A	31.1	24.9	0.0	0.7	0.0
ImageNet-R	68.3	63.2	0.0	0.7	0.0
ImageNet-Sketch	42.5	41.1	0.0	0.7	0.0
ImageNetV2	57.2	48.9	0.0	0.7	0.0

Backbone: ViT-B/16 Label: loop

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	28.9	0.0	0.7	0.0
Cars	66.2	66.2	0.0	0.7	0.0
Flower102	73.0	72.0	0.0	0.7	0.0
Food101	84.2	82.5	0.0	0.8	0.0
Pets	90.3	90.3	0.0	0.7	0.0
Caltech101	94.2	94.4	0.0	0.7	0.0
DTD	53.2	53.0	0.0	0.7	0.0
EuroSAT	55.8	53.0	1.0	0.7	0.0
SUN397	67.9	68.3	0.0	0.7	0.0
UCF101	67.0	67.4	0.0	0.7	0.0
ImageNet	69.7	70.2	0.0	0.7	0.0
ImageNet-A	50.7	49.1	0.0	0.7	0.0
ImageNet-R	77.1	79.0	0.0	0.7	0.0
ImageNet-Sketch	49.1	49.9	0.0	0.7	0.0
ImageNetV2	63.4	63.3	0.0	0.7	0.0

Backbone: ViT-B/32 Label: loop

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	23.1	0.0	0.7	0.0
Cars	61.2	63.2	0.0	0.7	0.0
Flower102	67.0	68.1	0.0	0.7	0.0
Food101	78.7	79.2	0.0	0.7	0.0
Pets	86.6	86.8	0.0	0.7	0.0
Caltech101	93.3	93.3	0.0	0.6	0.0
DTD	49.9	51.2	0.0	0.7	0.0
EuroSAT	50.8	49.6	1.0	0.6	0.0
SUN397	65.7	66.9	0.0	0.7	0.0
UCF101	63.6	65.8	0.0	0.7	0.0
ImageNet	64.4	66.0	0.0	0.7	0.0
ImageNet-A	31.1	31.5	0.0	0.7	0.0
ImageNet-R	68.3	71.5	0.0	0.7	0.0
ImageNet-Sketch	42.5	44.1	0.0	0.6	0.0
ImageNetV2	57.2	58.2	0.0	0.7	0.0

Backbone: ViT-B/16 Label: fixed

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	28.7	0.0	0.7	0.0
Cars	66.2	65.4	0.0	0.7	0.0
Flower102	73.0	71.9	0.0	0.7	0.0
Food101	84.2	82.4	0.0	0.8	0.0
Pets	90.3	90.4	0.0	0.7	0.0
Caltech101	94.2	94.3	0.0	0.7	0.0
DTD	53.2	53.1	0.0	0.7	0.0
EuroSAT	55.8	53.1	0.6	0.7	0.0
SUN397	67.9	68.1	0.0	0.7	0.0
UCF101	67.0	67.2	0.0	0.7	0.0
ImageNet	69.7	70.0	0.0	0.7	0.0
ImageNet-A	50.7	49.0	0.0	0.7	0.0
ImageNet-R	77.1	78.7	0.0	0.7	0.0
ImageNet-Sketch	49.1	49.7	0.0	0.7	0.0
ImageNetV2	63.4	63.2	0.0	0.7	0.0

Backbone: ViT-B/32 Label: fixed

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	22.2	0.0	0.7	0.0
Cars	61.2	62.1	0.0	0.7	0.0
Flower102	67.0	68.2	0.0	0.7	0.0
Food101	78.7	79.2	0.0	0.7	0.0
Pets	86.6	86.8	0.0	0.7	0.0
Caltech101	93.3	93.2	0.0	0.6	0.0
DTD	49.9	51.2	0.0	0.7	0.0
EuroSAT	50.8	53.4	0.8	0.7	0.0
SUN397	65.7	66.5	0.0	0.7	0.0
UCF101	63.6	65.1	0.0	0.7	0.0
ImageNet	64.4	65.6	0.0	0.7	0.0
ImageNet-A	31.1	31.2	0.0	0.7	0.0
ImageNet-R	68.3	71.2	0.0	0.7	0.0
ImageNet-Sketch	42.5	43.9	0.0	0.6	0.0
ImageNetV2	57.2	58.1	0.0	0.7	0.0

Table 10: Dataset-level RQ3 label-ablation results under the AutoAttack attack setting from all currently available raw data (all-sample scope). Clean and DBD are reported as percentages.

Backbone: RN50 Label: GT

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	22.5	5.9	0.8	0.0
Cars	57.3	60.7	10.8	0.8	0.0
Flower102	65.9	67.2	11.2	0.9	0.1
Food101	75.6	77.9	28.4	0.9	0.1
Pets	85.0	85.8	10.9	0.9	0.1
Caltech101	89.2	90.2	8.4	0.9	0.1
DTD	48.4	51.0	10.8	0.8	0.0
EuroSAT	37.6	42.2	12.5	0.7	0.0
SUN397	62.8	66.2	5.9	0.9	0.0
UCF101	59.7	62.8	14.1	0.8	0.1
ImageNet	61.5	65.2	3.8	0.8	0.0
ImageNet-A	23.9	26.2	6.3	0.8	0.0
ImageNet-R	60.1	65.2	7.0	0.8	0.1
ImageNet-Sketch	35.6	39.8	0.8	0.7	0.0
ImageNetV2	54.6	58.2	3.8	0.8	0.0

Backbone: ViT-B/16 Label: GT

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	30.5	8.3	0.8	0.0
Cars	66.4	67.8	20.8	0.8	0.1
Flower102	73.0	73.9	17.5	0.9	0.1
Food101	84.2	84.9	35.2	0.9	0.1
Pets	90.3	89.8	11.9	0.9	0.1
Caltech101	94.2	94.4	11.2	0.9	0.1
DTD	53.2	55.0	21.5	0.8	0.0
EuroSAT	55.8	47.4	34.2	0.8	0.0
SUN397	66.5	69.3	11.2	0.9	0.1
UCF101	67.0	68.6	17.6	0.9	0.1
ImageNet	69.8	71.9	5.9	0.9	0.1
ImageNet-A	50.7	53.4	17.2	0.8	0.0
ImageNet-R	77.1	80.4	11.8	0.9	0.1
ImageNet-Sketch	48.0	51.4	2.3	0.8	0.0
ImageNetV2	63.4	65.6	6.3	0.8	0.1

Backbone: RN50 Label: pseudo

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	18.3	6.8	0.9	-0.1
Cars	57.3	54.9	11.9	0.9	0.0
Flower102	65.9	64.7	10.9	0.9	0.0
Food101	75.6	74.7	28.5	0.9	0.0
Pets	85.0	82.9	10.5	0.9	0.0
Caltech101	89.2	88.1	9.2	0.9	0.1
DTD	48.4	46.9	11.6	0.9	0.0
EuroSAT	37.6	32.8	12.5	0.9	-0.1
SUN397	62.8	62.1	5.9	0.9	0.0
UCF101	59.7	58.2	14.3	0.9	0.0
ImageNet	61.5	60.6	3.9	0.9	0.0
ImageNet-A	23.9	22.8	7.0	0.9	-0.2
ImageNet-R	60.1	59.8	7.0	0.9	0.0
ImageNet-Sketch	35.6	35.4	1.0	0.9	-0.1
ImageNetV2	54.6	53.9	3.9	0.9	-0.1

Backbone: ViT-B/16 Label: pseudo

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	25.8	9.5	0.9	-0.1
Cars	66.2	64.5	21.4	0.9	0.0
Flower102	73.0	72.0	18.4	0.9	0.0
Food101	84.2	82.6	35.2	0.9	0.1
Pets	90.3	88.5	11.8	0.9	0.1
Caltech101	94.2	93.6	12.3	0.9	0.1
DTD	53.2	52.0	21.8	0.9	0.0
EuroSAT	55.8	45.5	35.3	0.9	0.0
SUN397	67.9	66.9	9.3	0.9	0.0
UCF101	67.0	65.3	18.0	0.9	0.0
ImageNet	69.7	68.5	5.7	0.9	0.0
ImageNet-A	50.7	48.2	17.6	0.9	-0.1
ImageNet-R	77.1	76.6	11.6	0.9	0.0
ImageNet-Sketch	49.1	48.6	2.3	0.9	-0.1
ImageNetV2	63.4	62.2	6.3	0.9	0.0

Backbone: RN50 Label: loop

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	19.6	0.2	0.7	0.0
Cars	57.3	57.3	0.2	0.7	0.0
Flower102	65.9	56.5	0.0	0.8	-0.1
Food101	75.6	66.1	0.0	0.8	-0.1
Pets	85.0	84.1	0.0	0.8	0.0
Caltech101	89.2	88.6	0.0	0.7	0.0
DTD	48.4	45.3	0.0	0.8	0.0
EuroSAT	37.6	36.1	2.2	0.7	0.0
SUN397	62.8	61.1	0.0	0.7	0.0
UCF101	59.7	59.4	0.1	0.7	0.0
ImageNet	61.5	59.9	0.0	0.7	0.0
ImageNet-A	23.9	19.0	0.0	0.7	0.0
ImageNet-R	60.1	61.2	0.0	0.7	0.0
ImageNet-Sketch	35.6	36.2	0.0	0.7	0.0
ImageNetV2	54.6	52.0	0.0	0.7	0.0

Backbone: ViT-B/16 Label: loop

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	28.2	0.2	0.7	0.0
Cars	66.2	65.8	0.5	0.7	0.0
Flower102	73.0	72.1	0.0	0.7	0.0
Food101	74.2	71.5	0.1	0.8	0.0
Pets	90.3	90.4	0.0	0.7	0.0
Caltech101	94.2	94.1	0.0	0.7	0.0
DTD	53.2	52.5	0.0	0.7	0.0
EuroSAT	55.8	51.4	0.7	0.7	0.0
SUN397	67.9	68.0	0.0	0.7	0.0
UCF101	67.0	66.6	0.1	0.7	0.0
ImageNet	69.7	69.9	0.0	0.7	0.0
ImageNet-A	50.7	48.5	0.0	0.7	0.0
ImageNet-R	77.1	78.4	0.0	0.7	0.0
ImageNet-Sketch	49.1	49.6	0.0	0.7	0.0
ImageNetV2	63.4	63.1	0.0	0.7	0.0

Backbone: RN50 Label: fixed

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	19.4	0.0	0.7	0.0
Cars	57.3	57.4	0.0	0.7	0.0
Flower102	65.9	56.5	0.0	0.8	-0.1
Food101	75.6	66.1	0.0	0.8	-0.1
Pets	85.0	84.2	0.0	0.8	0.0
Caltech101	89.2	88.6	0.0	0.7	0.0
DTD	48.4	45.4	0.0	0.8	0.0
EuroSAT	37.6	35.7	2.0	0.7	0.0
SUN397	62.8	61.2	0.0	0.7	0.0
UCF101	59.7	59.5	0.0	0.7	0.0
ImageNet	61.5	60.1	0.0	0.7	0.0
ImageNet-A	23.9	19.2	0.0	0.7	0.0
ImageNet-R	60.1	61.4	0.0	0.7	0.0
ImageNet-Sketch	35.6	36.3	0.0	0.7	0.0
ImageNetV2	54.6	52.1	0.0	0.7	0.0

Backbone: ViT-B/16 Label: fixed

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	28.1	0.0	0.7	0.0
Cars	66.2	65.5	0.0	0.7	0.0
Flower102	73.0	72.0	0.0	0.7	0.0
Food101	75.3	72.6	0.0	0.8	0.0
Pets	90.3	90.3	0.0	0.7	0.0
Caltech101	94.2	94.2	0.0	0.7	0.0
DTD	53.2	52.3	0.0	0.7	0.0
EuroSAT	55.8	51.7	0.6	0.7	0.0
SUN397	66.6	66.1	0.0	0.7	0.0
UCF101	67.0	67.0	0.0	0.7	0.0
ImageNet	69.7	70.0	0.0	0.7	0.0
ImageNet-A	50.7	48.9	0.0	0.7	0.0
ImageNet-R	77.1	78.6	0.0	0.7	0.0
ImageNet-Sketch	49.1	49.7	0.0	0.7	0.0
ImageNetV2	63.4	63.2	0.0	0.7	0.0

Backbone: ViT-B/32 Label: GT

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	21.5	5.5	0.7	0.0
Cars	61.2	61.9	8.6	0.8	0.0
Flower102	67.0	64.7	8.0	0.8	0.1
Food101	78.7	75.1	20.2	0.8	0.1
Pets	86.6	83.3	9.2	0.8	0.1
Caltech101	93.3	89.2	7.6	0.9	0.1
DTD	49.9	49.0	14.5	0.8	0.0
EuroSAT	50.8	39.1	27.6	0.8	0.0
SUN397	65.7	64.6	4.2	0.8	0.0
UCF101	63.6	61.6	10.9	0.8	0.0
ImageNet	64.4	63.2	2.9	0.8	0.0
ImageNet-A	31.1	28.9	8.1	0.7	0.0
ImageNet-R	68.3	68.0	6.2	0.8	0.1
ImageNet-Sketch	42.5	42.8	1.0	0.7	0.0
ImageNetV2	57.2	56.0	2.8	0.8	0.0

Backbone: ViT-B/32 Label: loop

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	22.2	0.4	0.7	0.0
Cars	60.6	62.7	0.2	0.7	0.0
Flower102	67.0	67.8	0.0	0.7	0.0
Food101	78.7	79.2	0.0	0.7	0.0
Pets	86.6	86.5	0.0	0.7	0.0
Caltech101	93.3	93.0	0.0	0.6	0.0
DTD	49.9	51.5	0.0	0.7	0.0
EuroSAT	44.0	43.3	0.0	0.6	0.0
SUN397	65.7	66.3	0.0	0.7	0.0
UCF101	63.6	64.8	0.1	0.7	0.0
ImageNet	64.9	65.9	0.0	0.7	0.0
ImageNet-A	31.1	30.7	0.0	0.7	0.0
ImageNet-R	68.3	70.9	0.0	0.7	0.0
ImageNet-Sketch	41.0	42.3	0.0	0.6	0.0
ImageNetV2	57.2	58.2	0.0	0.7	0.0

Backbone: ViT-B/32 Label: pseudo

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	20.0	7.3	0.9	-0.1
Cars	61.1	58.3	8.6	0.9	0.0
Flower102	67.0	61.9	7.5	0.9	0.0
Food101	76.4	69.5	34.1	0.9	0.1
Pets	86.6	81.6	9.1	0.9	0.0
Caltech101	93.3	88.7	7.4	0.9	0.1
DTD	49.9	46.0	15.3	0.9	0.0
EuroSAT	40.4	34.1	23.8	0.9	0.0
SUN397	63.8	60.1	4.0	0.9	0.0
UCF101	63.6	58.5	11.0	0.9	0.0
ImageNet	64.4	59.9	2.9	0.9	0.0
ImageNet-A	31.1	26.3	8.0	0.9	-0.1
ImageNet-R	68.3	64.5	6.3	0.9	0.0
ImageNet-Sketch	42.5	40.1	1.1	0.9	-0.1
ImageNetV2	57.2	52.4	3.1	0.9	0.0

Backbone: ViT-B/32 Label: fixed

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	22.0	0.0	0.7	0.0
Cars	61.2	62.0	0.0	0.7	0.0
Flower102	67.0	67.8	0.0	0.7	0.0
Food101	78.7	79.2	0.0	0.7	0.0
Pets	86.6	86.6	0.0	0.7	0.0
Caltech101	93.3	93.1	0.0	0.6	0.0
DTD	49.9	51.4	0.0	0.7	0.0
EuroSAT	50.8	49.2	0.8	0.7	0.0
SUN397	65.7	66.5	0.0	0.7	0.0
UCF101	63.6	65.0	0.0	0.7	0.0
ImageNet	67.3	68.6	0.0	0.7	0.0
ImageNet-A	31.1	30.9	0.0	0.7	0.0
ImageNet-R	67.6	70.0	0.0	0.7	0.0
ImageNet-Sketch	32.9	33.6	0.0	0.6	0.0
ImageNetV2	57.2	58.3	0.0	0.7	0.0

Table 11: Dataset-level RQ3 label-ablation results under the PGD-10 attack setting from all currently available raw data (all-sample scope). Clean and DBD are reported as percentages.

Backbone: RN50 Label: GT

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	60.7	27.5	0.9	0.0
Cars	57.3	88.7	29.5	0.9	0.1
Flower102	66.0	88.4	27.9	0.9	0.1
Food101	75.5	96.6	50.3	0.9	0.2
Pets	85.0	95.3	17.9	0.9	0.1
Caltech101	89.2	97.6	16.5	0.9	0.1
DTD	48.5	85.8	27.9	0.9	0.1
EuroSAT	37.6	76.1	54.9	0.9	0.0
SUN397	62.8	92.8	15.2	0.9	0.1
UCF101	59.7	87.9	32.6	0.9	0.1
ImageNet	61.5	94.2	12.5	0.9	0.1
ImageNet-A	23.9	89.1	42.2	0.9	0.1
ImageNet-R	60.0	96.2	25.3	0.9	0.1
ImageNet-Sketch	35.6	84.9	9.1	0.9	0.1
ImageNetV2	54.6	92.9	14.1	0.9	0.1

Backbone: RN50 Label: loop

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	7.6	2.4	0.9	-0.1
Cars	57.3	5.6	0.6	0.9	-0.3
Flower102	66.0	2.1	0.0	0.9	-0.4
Food101	75.5	3.6	0.0	0.9	-0.4
Pets	85.0	0.3	0.0	0.9	-0.5
Caltech101	89.2	8.1	0.0	0.9	-0.3
DTD	48.5	3.4	0.1	0.9	-0.2
EuroSAT	37.6	7.8	0.9	0.9	-0.2
SUN397	62.8	8.1	0.0	0.9	-0.3
UCF101	59.7	8.5	0.3	0.9	-0.3
ImageNet	61.5	4.9	0.0	0.9	-0.3
ImageNet-A	23.9	5.3	0.2	0.9	-0.2
ImageNet-R	60.0	5.3	0.0	0.9	-0.3
ImageNet-Sketch	35.6	7.4	0.2	0.9	-0.2
ImageNetV2	54.6	4.4	0.0	0.9	-0.3

Backbone: RN50 Label: pseudo

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	17.5	8.5	0.9	-0.1
Cars	57.3	55.5	18.4	0.9	0.0
Flower102	66.0	65.1	19.2	0.9	0.0
Food101	75.5	75.2	37.8	0.9	0.1
Pets	85.0	82.4	15.7	0.9	0.1
Caltech101	89.2	88.3	14.0	0.9	0.1
DTD	48.5	47.4	16.2	0.9	-0.1
EuroSAT	37.6	36.3	20.8	0.9	-0.1
SUN397	62.8	62.2	10.4	0.9	0.0
UCF101	59.7	58.9	21.4	0.9	0.0
ImageNet	61.5	60.7	7.0	0.9	0.0
ImageNet-A	23.9	23.4	9.3	0.9	-0.2
ImageNet-R	60.0	59.8	11.8	0.9	0.0
ImageNet-Sketch	35.6	35.4	1.9	0.9	-0.1
ImageNetV2	54.6	54.0	6.7	0.9	-0.1

Backbone: RN50 Label: fixed

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	1.2	0.0	0.9	-0.3
Cars	57.3	1.2	0.0	0.9	-0.5
Flower102	66.0	0.7	0.0	0.9	-0.4
Food101	75.5	2.3	0.0	0.9	-0.4
Pets	85.0	1.0	0.0	0.9	-0.5
Caltech101	89.2	16.0	0.0	0.9	-0.3
DTD	48.5	2.3	0.0	0.9	-0.2
EuroSAT	37.6	3.2	0.1	0.9	-0.2
SUN397	62.8	6.7	0.0	0.9	-0.3
UCF101	59.7	15.6	0.0	0.9	-0.3
ImageNet	61.5	2.5	0.0	0.9	-0.4
ImageNet-A	23.9	2.1	0.0	0.9	-0.4
ImageNet-R	60.0	5.0	0.0	0.9	-0.4
ImageNet-Sketch	35.6	4.5	0.0	0.9	-0.3
ImageNetV2	54.6	2.4	0.0	0.9	-0.4

Backbone: ViT-B/16 Label: GT

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	64.7	26.1	0.9	0.0
Cars	66.2	90.3	38.0	0.9	0.1
Flower102	73.0	91.1	30.8	0.9	0.1
Food101	84.2	96.7	47.6	0.9	0.2
Pets	90.3	96.2	15.1	0.9	0.1
Caltech101	94.2	99.0	16.0	0.9	0.2
DTD	53.2	83.9	33.9	0.9	0.1
EuroSAT	55.8	71.1	72.9	0.9	0.1
SUN397	67.9	93.8	17.2	0.9	0.1
UCF101	67.0	91.1	33.5	0.9	0.1
ImageNet	69.7	94.6	12.5	0.9	0.1
ImageNet-A	50.7	90.9	42.9	0.9	0.1
ImageNet-R	77.1	97.6	23.9	0.9	0.1
ImageNet-Sketch	49.1	90.2	10.0	0.9	0.1
ImageNetV2	63.4	93.3	15.5	0.9	0.1

Backbone: ViT-B/32 Label: GT

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	54.3	23.6	0.9	0.0
Cars	61.2	84.3	18.5	0.9	0.1
Flower102	67.0	81.9	14.9	0.9	0.1
Food101	78.7	90.0	28.1	0.9	0.1
Pets	86.6	94.7	10.9	0.9	0.1
Caltech101	93.3	95.7	8.8	0.9	0.1
DTD	49.9	74.1	22.8	0.9	0.0
EuroSAT	50.8	56.1	69.9	0.9	0.0
SUN397	65.7	85.2	6.8	0.9	0.1
UCF101	63.6	80.8	22.0	0.9	0.1
ImageNet	64.4	85.7	5.8	0.9	0.1
ImageNet-A	31.1	68.6	31.2	0.9	0.1
ImageNet-R	68.3	88.8	14.3	0.9	0.1
ImageNet-Sketch	42.5	77.3	6.4	0.9	0.0
ImageNetV2	57.2	82.3	7.2	0.9	0.1

Backbone: ViT-B/16 Label: pseudo

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	26.3	11.7	0.9	-0.1
Cars	66.2	64.8	26.9	0.9	0.0
Flower102	73.0	71.7	23.2	0.9	0.0
Food101	84.2	83.4	40.4	0.9	0.1
Pets	90.3	88.2	13.8	0.9	0.1
Caltech101	94.2	93.8	14.9	0.9	0.1
DTD	53.2	52.5	23.5	0.9	0.0
EuroSAT	55.8	47.7	42.0	0.9	0.0
SUN397	67.9	67.3	12.9	0.9	0.0
UCF101	67.0	65.6	23.6	0.9	0.0
ImageNet	69.7	68.9	8.0	0.9	0.0
ImageNet-A	50.7	49.2	21.4	0.9	-0.1
ImageNet-R	77.1	76.7	15.8	0.9	0.0
ImageNet-Sketch	49.1	48.8	3.2	0.9	-0.1
ImageNetV2	63.4	62.8	8.8	0.9	0.0

Backbone: ViT-B/32 Label: pseudo

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	19.9	9.1	0.9	-0.1
Cars	61.2	59.3	11.4	0.9	0.0
Flower102	67.0	63.8	8.9	0.9	0.0
Food101	78.7	75.5	21.6	0.9	0.0
Pets	86.6	84.0	9.4	0.9	0.0
Caltech101	93.3	90.9	8.1	0.9	0.1
DTD	49.9	47.5	14.3	0.9	0.0
EuroSAT	50.8	44.7	36.2	0.9	0.0
SUN397	65.7	63.6	4.9	0.9	0.0
UCF101	63.6	60.8	14.9	0.9	0.0
ImageNet	64.4	61.6	3.1	0.9	0.0
ImageNet-A	31.1	28.7	9.2	0.9	-0.1
ImageNet-R	68.3	66.3	7.2	0.9	0.0
ImageNet-Sketch	42.5	41.4	1.5	0.9	-0.1
ImageNetV2	57.2	54.4	3.2	0.9	0.0

Backbone: ViT-B/16 Label: loop

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	12.0	3.2	0.9	-0.2
Cars	66.2	8.1	0.6	0.9	-0.3
Flower102	73.0	5.5	0.0	0.9	-0.4
Food101	84.2	10.9	0.0	0.9	-0.4
Pets	90.3	2.8	0.0	0.9	-0.4
Caltech101	94.2	16.7	0.0	0.9	-0.3
DTD	53.2	7.6	0.0	0.9	-0.2
EuroSAT	55.8	9.6	0.4	0.9	-0.2
SUN397	67.9	12.8	0.1	0.9	-0.3
UCF101	67.0	11.6	0.1	0.9	-0.3
ImageNet	69.7	9.3	0.0	0.9	-0.3
ImageNet-A	50.7	9.9	0.1	0.9	-0.3
ImageNet-R	77.1	8.0	0.1	0.9	-0.3
ImageNet-Sketch	49.1	9.5	0.2	0.9	-0.2
ImageNetV2	63.4	8.7	0.0	0.9	-0.3

Backbone: ViT-B/32 Label: loop

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	12.8	2.9	0.9	-0.1
Cars	61.2	21.3	0.5	0.9	-0.2
Flower102	67.0	20.2	0.0	0.9	-0.2
Food101	78.7	28.3	0.0	0.9	-0.3
Pets	86.6	17.8	0.0	0.9	-0.3
Caltech101	93.3	55.6	0.0	0.9	-0.2
DTD	49.9	21.4	0.1	0.9	-0.1
EuroSAT	50.8	21.4	0.9	0.9	-0.1
SUN397	65.7	33.2	0.0	0.9	-0.2
UCF101	63.6	26.5	0.2	0.9	-0.2
ImageNet	64.4	24.7	0.0	0.9	-0.2
ImageNet-A	31.1	14.1	0.2	0.9	-0.2
ImageNet-R	68.3	24.9	0.0	0.9	-0.2
ImageNet-Sketch	42.5	18.5	0.1	0.9	-0.2
ImageNetV2	57.2	22.9	0.0	0.9	-0.2

Backbone: ViT-B/16 Label: fixed

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	2.3	0.0	0.9	-0.2
Cars	66.2	5.0	0.0	0.9	-0.5
Flower102	73.0	2.9	0.0	0.9	-0.4
Food101	84.2	12.2	0.0	0.9	-0.3
Pets	90.3	4.3	0.0	0.9	-0.5
Caltech101	94.2	26.1	0.0	0.9	-0.3
DTD	53.2	3.6	0.0	0.9	-0.2
EuroSAT	55.8	8.3	0.1	0.9	-0.2
SUN397	67.9	13.0	0.0	0.9	-0.3
UCF101	67.0	26.1	0.0	0.9	-0.2
ImageNet	69.7	8.3	0.0	0.9	-0.4
ImageNet-A	50.7	8.2	0.0	0.9	-0.4
ImageNet-R	77.1	11.0	0.0	0.9	-0.3
ImageNet-Sketch	49.1	8.5	0.0	0.9	-0.3
ImageNetV2	63.4	7.3	0.0	0.9	-0.4

Backbone: ViT-B/32 Label: fixed

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	4.4	0.1	0.9	-0.2
Cars	61.2	18.6	0.0	0.8	-0.3
Flower102	67.0	15.1	0.0	0.8	-0.3
Food101	78.7	30.9	0.0	0.9	-0.2
Pets	86.6	17.7	0.0	0.9	-0.3
Caltech101	93.3	66.1	0.0	0.8	-0.2
DTD	49.9	18.6	0.1	0.9	-0.1
EuroSAT	50.8	19.6	0.9	0.9	-0.1
SUN397	65.7	32.0	0.0	0.9	-0.2
UCF101	63.6	42.4	0.0	0.8	-0.2
ImageNet	64.4	26.7	0.0	0.9	-0.3
ImageNet-A	31.1	12.9	0.0	0.9	-0.3
ImageNet-R	68.3	28.9	0.0	0.8	-0.2
ImageNet-Sketch	42.5	18.3	0.0	0.9	-0.2
ImageNetV2	57.2	23.3	0.0	0.9	-0.3

Table 12: Dataset-level RQ3 label-ablation results under the PGD-100 attack setting from all currently available raw data (all-sample scope). Clean and DBD are reported as percentages.

Backbone: RN50 Label: GT

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	70.2	52.5	0.9	0.1
Cars	57.3	91.4	65.4	0.9	0.1
Flower102	66.0	93.0	65.9	1.0	0.2
Food101	75.5	98.8	82.9	1.0	0.3
Pets	85.0	94.2	57.1	1.0	0.2
Caltech101	89.2	99.2	45.2	0.9	0.2
DTD	48.5	88.5	54.4	1.0	0.1
EuroSAT	37.6	84.3	82.5	0.9	0.1
SUN397	62.8	96.3	53.1	1.0	0.2
UCF101	59.7	92.4	63.9	0.9	0.2
ImageNet	61.5	96.3	49.8	1.0	0.2
ImageNet-A	23.9	97.6	79.0	0.9	0.2
ImageNet-R	60.0	98.7	65.7	0.9	0.2
ImageNet-Sketch	35.6	90.6	31.3	0.9	0.1
ImageNetV2	54.6	95.8	51.8	0.9	0.2

Backbone: ViT-B/16 Label: GT

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	71.4	52.5	1.0	0.1
Cars	66.4	94.0	72.2	1.0	0.2
Flower102	73.0	97.2	69.7	1.0	0.3
Food101	82.8	100.0	95.5	1.0	0.3
Pets	90.3	97.1	56.6	1.0	0.3
Caltech101	94.2	99.4	56.4	1.0	0.3
DTD	53.2	91.4	65.3	1.0	0.1
EuroSAT	45.0	100.0	99.9	1.0	0.1
SUN397	66.7	97.8	64.1	1.0	0.2
UCF101	67.0	95.8	74.2	1.0	0.2
ImageNet	69.7	97.7	56.6	1.0	0.2
ImageNet-A	50.7	98.9	82.3	0.9	0.3
ImageNet-R	77.1	99.4	66.4	1.0	0.3
ImageNet-Sketch	49.1	95.2	37.7	1.0	0.1
ImageNetV2	63.4	97.5	59.4	1.0	0.2

Backbone: RN50 Label: pseudo

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	18.5	14.8	0.9	-0.2
Cars	57.3	56.0	40.7	0.9	0.0
Flower102	66.0	65.2	48.7	1.0	0.0
Food101	75.5	75.4	62.9	1.0	0.1
Pets	85.0	81.7	51.0	1.0	0.1
Caltech101	89.2	88.9	39.1	0.9	0.2
DTD	48.5	47.2	32.0	1.0	-0.1
EuroSAT	37.6	36.9	34.4	0.9	-0.1
SUN397	62.8	62.4	36.4	1.0	0.0
UCF101	59.7	58.7	41.7	0.9	0.0
ImageNet	61.5	60.8	31.0	1.0	0.0
ImageNet-A	23.9	23.7	18.4	0.9	-0.3
ImageNet-R	60.0	59.9	36.9	0.9	0.0
ImageNet-Sketch	29.1	28.6	11.9	0.9	-0.2
ImageNetV2	54.6	54.2	28.7	0.9	-0.1

Backbone: ViT-B/16 Label: pseudo

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	26.5	20.3	1.0	-0.2
Cars	66.2	65.3	50.7	1.0	0.1
Flower102	73.0	72.5	52.5	1.0	0.1
Food101	84.2	84.1	71.4	1.0	0.2
Pets	90.3	88.9	51.7	1.0	0.2
Caltech101	94.2	94.0	53.4	1.0	0.3
DTD	53.2	52.8	41.9	1.0	-0.1
EuroSAT	55.8	53.9	52.1	1.0	0.0
SUN397	67.9	67.5	45.1	1.0	0.0
UCF101	67.0	66.2	51.7	1.0	0.0
ImageNet	70.3	69.9	39.9	1.0	0.0
ImageNet-A	54.9	54.6	45.9	1.0	-0.1
ImageNet-R	77.1	77.0	49.5	1.0	0.1
ImageNet-Sketch	47.7	47.5	16.6	1.0	-0.1
ImageNetV2	63.7	62.7	37.4	1.0	0.0

Backbone: RN50 Label: loop

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	6.6	4.1	0.9	-0.2
Cars	57.3	3.9	1.3	0.9	-0.5
Flower102	66.0	0.2	0.0	1.0	-0.7
Food101	75.5	0.6	0.0	1.0	-0.7
Pets	85.0	0.1	0.0	1.0	-0.8
Caltech101	89.2	0.7	0.0	0.9	-0.6
DTD	48.5	1.3	0.0	1.0	-0.4
EuroSAT	37.6	1.3	0.3	0.9	-0.3
SUN397	62.8	2.6	0.1	0.9	-0.5
UCF101	59.7	3.2	0.5	0.9	-0.4
ImageNet	61.5	1.6	0.1	0.9	-0.6
ImageNet-A	23.9	1.2	0.1	0.9	-0.4
ImageNet-R	60.0	1.0	0.0	0.9	-0.5
ImageNet-Sketch	30.6	3.9	1.6	0.9	-0.3
ImageNetV2	54.6	1.3	0.1	0.9	-0.6

Backbone: ViT-B/16 Label: loop

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	8.6	4.3	1.0	-0.2
Cars	66.2	3.0	0.8	0.9	-0.5
Flower102	73.0	0.5	0.0	1.0	-0.7
Food101	84.1	0.7	0.0	1.0	-0.6
Pets	90.3	0.1	0.0	1.0	-0.7
Caltech101	94.2	0.5	0.0	0.9	-0.6
DTD	53.2	1.6	0.0	1.0	-0.4
EuroSAT	55.8	0.4	0.0	1.0	-0.4
SUN397	67.9	2.4	0.2	1.0	-0.5
UCF101	67.0	2.0	0.1	0.9	-0.5
ImageNet	70.6	2.2	0.2	1.0	-0.5
ImageNet-A	56.3	1.2	0.0	0.9	-0.5
ImageNet-R	77.1	0.9	0.1	0.9	-0.6
ImageNet-Sketch	46.1	2.7	0.4	0.9	-0.4
ImageNetV2	63.4	1.5	0.1	1.0	-0.6

Backbone: RN50 Label: fixed

Dataset	Clean	DBD	P	ρ	G
Aircraft	19.5	0.2	0.1	0.9	-0.4
Cars	57.3	0.3	0.1	0.9	-0.7
Flower102	66.0	0.2	0.0	1.0	-0.8
Food101	75.5	0.1	0.0	1.0	-0.7
Pets	85.0	0.2	0.0	1.0	-0.8
Caltech101	89.2	3.1	0.0	0.9	-0.5
DTD	48.5	0.5	0.0	1.0	-0.4
EuroSAT	37.6	0.0	0.0	0.9	-0.3
SUN397	62.8	1.8	0.0	0.9	-0.5
UCF101	59.7	4.5	0.0	0.9	-0.5
ImageNet	61.5	0.1	0.0	0.9	-0.8
ImageNet-A	23.9	0.1	0.0	0.9	-0.7
ImageNet-R	60.0	0.5	0.0	0.9	-0.6
ImageNet-Sketch	35.6	0.6	0.0	0.9	-0.5
ImageNetV2	54.6	0.2	0.0	0.9	-0.7

Backbone: ViT-B/16 Label: fixed

Dataset	Clean	DBD	P	ρ	G
Aircraft	27.2	0.2	0.0	1.0	-0.4
Cars	66.2	0.4	0.0	0.9	-0.8
Flower102	73.0	0.2	0.0	1.0	-0.7
Food101	81.6	0.3	0.0	1.0	-0.6
Pets	90.3	0.5	0.0	1.0	-0.8
Caltech101	94.2	1.1	0.0	0.9	-0.5
DTD	53.2	0.6	0.0	1.0	-0.4
EuroSAT	55.8	0.2	0.0	1.0	-0.4
SUN397	67.5	1.8	0.0	1.0	-0.5
UCF101	67.0	3.2	0.0	0.9	-0.5
ImageNet	69.7	0.3	0.0	0.9	-0.7
ImageNet-A	50.7	0.2	0.0	0.9	-0.7
ImageNet-R	75.5	0.8	0.0	0.9	-0.6
ImageNet-Sketch	49.1	0.9	0.0	0.9	-0.5
ImageNetV2	63.4	0.3	0.0	0.9	-0.7

Backbone: ViT-B/32 Label: GT

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	65.9	44.0	0.9	0.0
Cars	61.2	91.1	50.1	0.9	0.1
Flower102	67.0	94.0	45.0	0.9	0.2
Food101	78.7	98.4	63.3	0.9	0.2
Pets	86.6	96.7	40.7	0.9	0.2
Caltech101	93.3	99.2	34.6	0.9	0.2
DTD	49.9	87.3	48.0	0.9	0.1
EuroSAT	50.8	92.8	93.4	0.9	0.1
SUN397	65.7	94.1	33.4	0.9	0.1
UCF101	63.6	92.4	56.0	0.9	0.1
ImageNet	64.4	95.4	32.1	0.9	0.1
ImageNet-A	31.1	95.3	66.2	0.9	0.2
ImageNet-R	68.3	98.1	47.0	0.9	0.2
ImageNet-Sketch	42.5	89.7	23.0	0.9	0.1
ImageNetV2	57.2	94.2	33.6	0.9	0.1

Backbone: ViT-B/32 Label: loop

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	8.7	5.0	0.9	-0.2
Cars	61.2	8.2	1.2	0.9	-0.3
Flower102	67.0	3.1	0.0	0.9	-0.4
Food101	78.7	2.4	0.0	0.9	-0.5
Pets	86.6	0.6	0.0	0.9	-0.6
Caltech101	93.3	3.4	0.0	0.9	-0.5
DTD	49.9	4.7	0.0	0.9	-0.3
EuroSAT	50.8	1.2	0.1	0.9	-0.3
SUN397	65.7	8.6	0.0	0.9	-0.3
UCF101	63.6	4.2	0.3	0.9	-0.4
ImageNet	64.4	5.2	0.1	0.9	-0.4
ImageNet-A	31.1	3.2	0.1	0.9	-0.3
ImageNet-R	68.3	3.4	0.1	0.9	-0.4
ImageNet-Sketch	43.2	6.5	2.2	0.9	-0.2
ImageNetV2	57.2	4.6	0.1	0.9	-0.4

Backbone: ViT-B/32 Label: pseudo

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	19.7	14.4	0.9	-0.1
Cars	61.2	60.4	33.2	0.9	0.0
Flower102	67.0	66.2	30.8	0.9	0.0
Food101	78.7	78.5	49.8	0.9	0.1
Pets	86.6	84.7	36.1	0.9	0.1
Caltech101	93.3	93.1	32.5	0.9	0.2
DTD	49.9	48.9	29.1	0.9	0.0
EuroSAT	50.8	50.7	46.8	0.9	-0.1
SUN397	65.7	65.2	24.2	0.9	0.0
UCF101	63.6	62.6	38.2	0.9	0.0
ImageNet	64.4	63.8	20.2	0.9	0.0
ImageNet-A	31.1	30.8	20.4	0.9	-0.2
ImageNet-R	68.3	68.1	28.8	0.9	0.0
ImageNet-Sketch	42.5	42.0	7.3	0.9	-0.1
ImageNetV2	57.2	56.6	18.4	0.9	-0.1

Backbone: ViT-B/32 Label: fixed

Dataset	Clean	DBD	P	ρ	G
Aircraft	20.6	0.5	0.1	0.9	-0.3
Cars	61.2	2.6	0.0	0.9	-0.5
Flower102	67.0	0.6	0.0	0.9	-0.5
Food101	78.7	3.1	0.0	0.9	-0.4
Pets	86.6	0.5	0.0	0.9	-0.6
Caltech101	93.3	6.3	0.0	0.9	-0.4
DTD	49.9	3.8	0.1	0.9	-0.2
EuroSAT	50.8	0.1	0.1	0.9	-0.3
SUN397	65.7	7.8	0.0	0.9	-0.3
UCF101	63.6	11.7	0.0	0.9	-0.3
ImageNet	64.4	2.0	0.0	0.9	-0.5
ImageNet-A	31.1	1.6	0.0	0.9	-0.5
ImageNet-R	68.3	3.2	0.0	0.9	-0.4
ImageNet-Sketch	42.5	3.6	0.0	0.9	-0.4
ImageNetV2	57.2	1.8	0.0	0.9	-0.5

B.5 Detailed Results for RQ4: CLIP vs. DBD vs. **D-Pred**

Table 13 collects per-dataset clean and robust accuracy for zero-shot CLIP, the full DBD defense (Liu et al., 2026), and **D-Pred** (Ours), which classifies using only the recovered direction $\bar{\mathbf{d}}$, across five attacks (FGSM, PGD-10, PGD-100, CW, AA) on all 15 datasets and three backbones. Each sub-table corresponds to a backbone–attack pair. Clean accuracy for CLIP and DBD follows DBD (Liu et al., 2026) Table 1; **D-Pred** (Ours) clean accuracy and all robust accuracies are from our experiments.

Table 13: Per-dataset clean and robust accuracy (%) of zero-shot CLIP, DBD, and **D-Pred** across five attacks.

Backbone: RN50 Attack: FGSM

Dataset	CLIP		DBD		**D-Pred** (Ours)	
	Clean	Robust	Clean	Robust	Clean	Robust
FGVC Aircraft	19.6	3.2	21.6	11.8	3.7	26.0
Stanford Cars	57.3	3.6	60.0	28.7	0.2	12.6
Flower102	65.9	9.1	65.9	38.3	0.4	48.8
Food101	75.6	7.6	75.0	35.5	1.4	56.0
Oxford Pets	85.0	12.2	86.0	60.5	2.4	54.1
Caltech101	89.1	51.1	90.1	74.8	2.8	42.5
DTD	48.5	33.9	47.9	37.8	5.0	51.5
EuroSAT	37.5	20.8	29.4	17.5	1.7	24.7
SUN397	62.7	15.7	63.8	43.7	1.0	40.1
UCF101	59.7	22.9	60.6	34.9	3.9	46.2
ImageNet	61.5	11.6	63.1	42.7	0.6	39.4
ImageNet-A	23.8	2.0	23.0	7.8	2.4	41.6
ImageNet-R	60.0	21.7	63.0	44.4	3.2	58.7
ImageNet-Sketch	35.6	25.0	38.1	28.1	1.1	26.3
ImageNetV2	54.7	11.2	55.7	34.3	0.5	36.2

Backbone: RN50 Attack: PGD-10

Dataset	CLIP		DBD		**D-Pred** (Ours)	
	Clean	Robust	Clean	Robust	Clean	Robust
FGVC Aircraft	19.6	0.0	21.6	60.7	3.7	70.3
Stanford Cars	57.3	0.0	60.0	88.7	0.2	86.6
Flower102	65.9	0.0	65.9	88.4	0.4	93.7
Food101	75.6	0.0	75.0	96.6	1.4	98.1
Oxford Pets	85.0	0.0	86.0	95.3	2.4	85.2
Caltech101	89.1	0.5	90.1	97.6	2.8	90.2
DTD	48.5	0.0	47.9	85.8	5.0	89.6
EuroSAT	37.5	0.0	29.4	76.1	1.7	84.6
SUN397	62.7	0.0	63.8	92.8	1.0	93.4
UCF101	59.7	0.0	60.6	87.9	3.9	94.6
ImageNet	61.5	0.0	63.1	94.2	0.6	91.7
ImageNet-A	23.8	0.0	23.0	89.1	2.4	94.8
ImageNet-R	60.0	0.0	63.0	96.2	3.2	96.1
ImageNet-Sketch	35.6	0.0	38.1	84.9	1.1	88.4
ImageNetV2	54.7	0.0	55.7	92.9	0.5	91.4

Backbone: RN50 Attack: PGD-100

Dataset	CLIP		DBD		**D-Pred** (Ours)	
	Clean	Robust	Clean	Robust	Clean	Robust
FGVC Aircraft	19.6	0.0	21.6	70.2	3.7	83.9
Stanford Cars	57.3	0.0	60.0	91.4	0.2	92.2
Flower102	65.9	0.0	65.9	93.0	0.4	96.3
Food101	75.6	0.0	75.0	98.8	1.4	99.4
Oxford Pets	85.0	0.0	86.0	94.2	2.4	87.7
Caltech101	89.1	0.1	90.1	99.2	2.8	95.5
DTD	48.5	0.0	47.9	88.5	5.0	93.4
EuroSAT	37.5	0.0	29.4	84.3	1.7	93.6
SUN397	62.7	0.0	63.8	96.3	1.0	97.1
UCF101	59.7	0.0	60.6	92.4	3.9	96.6
ImageNet	61.5	0.0	63.1	96.3	0.6	95.6
ImageNet-A	23.8	0.0	23.0	97.6	2.4	98.7
ImageNet-R	60.0	0.0	63.0	98.7	3.2	98.0
ImageNet-Sketch	35.6	0.0	38.1	90.6	1.1	92.3
ImageNetV2	54.7	0.0	55.7	95.8	0.5	95.2

Backbone: ViT-B/16 Attack: PGD-10

Dataset	CLIP		DBD		**D-Pred** (Ours)	
	Clean	Robust	Clean	Robust	Clean	Robust
FGVC Aircraft	27.1	0.0	29.7	64.7	2.1	79.9
Stanford Cars	66.2	0.0	67.8	90.3	0.4	89.7
Flower102	73.0	0.0	73.6	91.1	1.4	94.2
Food101	84.2	0.0	84.7	96.7	1.2	98.3
Oxford Pets	90.3	0.0	90.7	96.2	7.4	85.5
Caltech101	94.2	0.3	94.8	99.0	2.6	92.7
DTD	53.2	0.0	54.5	83.9	7.2	94.4
EuroSAT	55.7	0.0	44.9	71.1	3.1	85.5
SUN397	67.9	0.0	69.0	93.8	1.2	94.1
UCF101	67.0	0.0	68.2	91.1	4.2	95.0
ImageNet	69.6	0.0	71.1	94.6	0.7	92.5
ImageNet-A	50.6	0.0	52.1	90.9	3.4	95.5
ImageNet-R	77.1	0.1	79.5	97.6	4.2	96.1
ImageNet-Sketch	49.1	0.1	51.1	90.2	1.2	88.6
ImageNetV2	63.4	0.0	64.7	93.3	0.9	92.2

Backbone: RN50 Attack: CW

Dataset	CLIP		DBD		**D-Pred** (Ours)	
	Clean	Robust	Clean	Robust	Clean	Robust
FGVC Aircraft	19.6	0.3	21.6	17.3	3.7	20.2
Stanford Cars	57.3	0.3	60.0	49.1	0.2	5.2
Flower102	65.9	0.7	65.9	57.7	0.4	37.0
Food101	75.6	0.2	75.0	65.4	1.4	55.9
Oxford Pets	85.0	0.5	86.0	80.0	2.4	39.7
Caltech101	89.1	4.9	90.1	85.6	2.8	58.7
DTD	48.5	2.6	47.9	45.3	5.0	52.2
EuroSAT	37.5	2.8	29.4	25.5	1.7	37.7
SUN397	62.7	0.6	63.8	58.5	1.0	33.5
UCF101	59.7	1.2	60.6	51.4	3.9	49.9
ImageNet	61.5	0.7	63.1	57.0	0.6	31.2
ImageNet-A	23.8	0.2	23.0	16.3	2.4	43.5
ImageNet-R	60.0	2.3	63.0	57.8	3.2	53.7
ImageNet-Sketch	35.6	5.2	38.1	34.4	1.1	19.1
ImageNetV2	54.7	0.6	55.7	49.4	0.5	30.1

Backbone: ViT-B/16 Attack: PGD-100

Dataset	CLIP		DBD		**D-Pred** (Ours)	
	Clean	Robust	Clean	Robust	Clean	Robust
FGVC Aircraft	27.1	0.0	29.7	71.4	2.1	86.2
Stanford Cars	66.2	0.0	67.8	94.0	0.4	93.5
Flower102	73.0	0.0	73.6	97.2	1.4	97.7
Food101	84.2	0.0	84.7	100.0	1.2	99.6
Oxford Pets	90.3	0.0	90.7	97.1	7.4	92.0
Caltech101	94.2	0.0	94.8	99.4	2.6	98.5
DTD	53.2	0.0	54.5	91.4	7.2	97.6
EuroSAT	55.7	0.0	44.9	100.0	3.1	95.4
SUN397	67.9	0.0	69.0	97.8	1.2	97.2
UCF101	67.0	0.0	68.2	95.8	4.2	97.7
ImageNet	69.6	0.0	71.1	97.7	0.7	97.2
ImageNet-A	50.6	0.0	52.1	98.9	3.4	98.9
ImageNet-R	77.1	0.0	79.5	99.4	4.2	98.4
ImageNet-Sketch	49.1	0.0	51.1	95.2	1.2	94.3
ImageNetV2	63.4	0.0	64.7	97.5	0.9	97.1

Backbone: RN50 Attack: AA

Dataset	CLIP		DBD		**D-Pred** (Ours)	
	Clean	Robust	Clean	Robust	Clean	Robust
FGVC Aircraft	19.6	0.0	21.6	22.5	3.7	67.5
Stanford Cars	57.3	0.0	60.0	60.7	0.2	82.0
Flower102	65.9	0.0	65.9	67.2	0.4	90.9
Food101	75.6	0.0	75.0	77.9	1.4	97.3
Oxford Pets	85.0	0.0	86.0	85.8	2.4	84.8
Caltech101	89.1	0.2	90.1	90.2	2.8	89.2
DTD	48.5	0.2	47.9	51.0	5.0	88.3
EuroSAT	37.5	0.0	29.4	42.2	1.7	61.7
SUN397	62.7	0.0	63.8	66.2	1.0	91.5
UCF101	59.7	0.0	60.6	62.8	3.9	92.3
ImageNet	61.5	0.0	63.1	65.2	0.6	89.9
ImageNet-A	23.8	0.1	23.0	26.2	2.4	90.8
ImageNet-R	60.0	0.0	63.0	65.2	3.2	94.9
ImageNet-Sketch	35.6	0.1	38.1	39.8	1.1	85.7
ImageNetV2	54.7	0.0	55.7	58.2	0.5	89.1

Backbone: ViT-B/16 Attack: CW

Dataset	CLIP		DBD		**D-Pred** (Ours)	
	Clean	Robust	Clean	Robust	Clean	Robust
FGVC Aircraft	27.1	0.1	29.7	23.4	2.1	19.4
Stanford Cars	66.2	0.1	67.8	48.7	0.4	12.9
Flower102	73.0	0.2	73.6	59.9	1.4	40.2
Food101	84.2	0.1	84.7	70.8	1.2	67.9
Oxford Pets	90.3	0.1	90.7	80.7	7.4	57.5
Caltech101	94.2	3.2	94.8	90.1	2.6	61.7
DTD	53.2	1.7	54.5	48.0	7.2	62.8
EuroSAT	55.7	1.2	44.9	26.6	3.1	58.3
SUN397	67.9	0.2	69.0	60.0	1.2	39.7
UCF101	67.0	0.6	68.2	56.3	4.2	57.6
ImageNet	69.6	0.5	71.1	61.5	0.7	37.0
ImageNet-A	50.6	0.2	52.1	37.1	3.4	60.7
ImageNet-R	77.1	1.1	79.5	70.4	4.2	67.6
ImageNet-Sketch	49.1	2.1	51.1	45.1	1.2	29.5
ImageNetV2	63.4	0.6	64.7	54.1	0.9	36.8

Backbone: ViT-B/16 Attack: FGSM

Dataset	CLIP		DBD		**D-Pred** (Ours)	
	Clean	Robust	Clean	Robust	Clean	Robust
FGVC Aircraft	27.1	0.3	29.7	15.9	2.1	50.5
Stanford Cars	66.2	1.0	67.8	26.6	0.4	46.6
Flower102	73.0	2.5	73.6	40.5	1.4	71.3
Food101	84.2	5.1	84.7	37.2	1.2	68.9
Oxford Pets	90.3	5.6	90.7	50.0	7.4	78.2
Caltech101	94.2	46.1	94.8	76.1	2.6	52.2
DTD	53.2	22.7	54.5	36.8	7.2	65.7
EuroSAT	55.7	19.2	44.9	15.0	3.1	28.7
SUN397	67.9	10.5	69.0	42.5	1.2	55.3
UCF101	67.0	12.3	68.2	33.0	4.2	61.1
ImageNet	69.6	8.4	71.1	40.6	0.7	52.2
ImageNet-A	50.6	2.1	52.1	13.8	3.4	51.4
ImageNet-R	77.1	17.1	79.5	48.4	4.2	73.0
ImageNet-Sketch	49.1	15.7	51.1	33.0	1.2	47.5
ImageNetV2	63.4	7.5	64.7	33.6	0.9	50.3

Backbone: ViT-B/16 Attack: AA

Dataset	CLIP		DBD		**D-Pred** (Ours)	
	Clean	Robust	Clean	Robust	Clean	Robust
FGVC Aircraft	27.1	0.0	29.7	30.5	2.1	75.2
Stanford Cars	66.2	0.0	67.8	67.8	0.4	87.4
Flower102	73.0	0.0	73.6	73.9	1.4	93.2
Food101	84.2	0.0	84.7	84.9	1.2	97.5
Oxford Pets	90.3	0.0	90.7	89.8	7.4	84.0
Caltech101	94.2	0.1	94.8	94.4	2.6	90.4
DTD	53.2	0.0	54.5	55.0	7.2	92.9
EuroSAT	55.7	0.0	44.9	47.4	3.1	80.7
SUN397	67.9	0.0	69.0	69.3	1.2	92.9
UCF101	67.0	0.0	68.2	68.6	4.2	93.9
ImageNet	69.6	0.0	71.1	71.9	0.7	91.1
ImageNet-A	50.6	0.0	52.1	53.4	3.4	93.2
ImageNet-R	77.1	0.0	79.5	80.4	4.2	95.0
ImageNet-Sketch	49.1	0.0	51.1	51.4	1.2	85.9
ImageNetV2	63.4	0.0	64.7	65.6	0.9	90.8

Backbone: ViT-B/32 Attack: FGSM

Dataset	CLIP		DBD		**D-Pred** (Ours)	
	Clean	Robust	Clean	Robust	Clean	Robust
FGVC Aircraft	20.6	0.7	22.4	8.5	1.9	36.8
Stanford Cars	61.2	1.5	63.7	14.8	0.2	32.6
Flower102	67.0	5.4	68.7	28.4	0.6	56.4
Food101	78.7	6.3	80.1	25.4	2.0	63.4
Oxford Pets	86.6	8.3	86.8	37.5	3.7	64.5
Caltech101	93.3	50.7	93.8	71.2	3.2	49.2
DTD	49.9	25.2	51.5	29.7	7.3	60.1
EuroSAT	50.8	24.4	41.7	19.4	2.4	31.5
SUN397	65.7	14.5	67.1	32.9	1.1	45.2
UCF101	63.6	15.2	65.3	31.1	4.7	57.8
ImageNet	64.4	10.6	66.3	28.6	0.5	41.9
ImageNet-A	31.1	1.8	32.9	5.2	2.8	43.9
ImageNet-R	68.3	17.7	71.7	35.7	3.7	63.5
ImageNet-Sketch	42.5	16.6	45.1	23.4	0.9	42.8
ImageNetV2	57.2	10.0	59.1	23.4	0.6	40.0

Backbone: ViT-B/32 Attack: CW

Dataset	CLIP		DBD		**D-Pred** (Ours)	
	Clean	Robust	Clean	Robust	Clean	Robust
FGVC Aircraft	20.6	0.3	22.4	16.3	1.9	16.6
Stanford Cars	61.2	0.2	63.7	43.4	0.2	7.3
Flower102	67.0	1.0	68.7	55.7	0.6	34.2
Food101	78.7	0.2	80.1	58.4	2.0	56.7
Oxford Pets	86.6	0.3	86.8	71.0	3.7	44.2
Caltech101	93.3	3.0	93.8	86.5	3.2	60.3
DTD	49.9	1.7	51.5	41.1	7.3	57.0
EuroSAT	50.8	1.7	41.7	27.0	2.4	50.6
SUN397	65.7	0.7	67.1	52.7	1.1	33.3
UCF101	63.6	1.0	65.3	50.8	4.7	54.6
ImageNet	64.4	0.8	66.3	50.8	0.5	30.2
ImageNet-A	31.1	0.1	32.9	18.8	2.8	50.9
ImageNet-R	68.3	1.7	71.7	59.4	3.7	57.5
ImageNet-Sketch	42.5	3.5	45.1	38.0	0.9	22.4
ImageNetV2	57.2	0.7	59.1	43.6	0.6	28.5

Backbone: ViT-B/32 Attack: PGD-10

Dataset	CLIP		DBD		**D-Pred** (Ours)	
	Clean	Robust	Clean	Robust	Clean	Robust
FGVC Aircraft	20.6	0.0	22.4	54.3	1.9	68.5
Stanford Cars	61.2	0.0	63.7	84.3	0.2	87.1
Flower102	67.0	0.0	68.7	81.9	0.6	92.5
Food101	78.7	0.0	80.1	90.0	2.0	95.8
Oxford Pets	86.6	0.0	86.8	94.7	3.7	83.2
Caltech101	93.3	0.9	93.8	95.7	3.2	90.6
DTD	49.9	0.0	51.5	74.1	7.3	91.8
EuroSAT	50.8	0.0	41.7	56.1	2.4	95.4
SUN397	65.7	0.0	67.1	85.2	1.1	90.5
UCF101	63.6	0.0	65.3	80.8	4.7	92.2
ImageNet	64.4	0.0	66.3	85.7	0.5	88.3
ImageNet-A	31.1	0.0	32.9	68.6	2.8	91.6
ImageNet-R	68.3	0.2	71.7	88.8	3.7	94.8
ImageNet-Sketch	42.5	0.2	45.1	77.3	0.9	85.8
ImageNetV2	57.2	0.0	59.1	82.3	0.6	88.2

Backbone: ViT-B/32 Attack: AA

Dataset	CLIP		DBD		**D-Pred** (Ours)	
	Clean	Robust	Clean	Robust	Clean	Robust
FGVC Aircraft	20.6	0.0	22.4	21.5	1.9	65.0
Stanford Cars	61.2	0.0	63.7	61.9	0.2	84.6
Flower102	67.0	0.0	68.7	64.7	0.6	91.4
Food101	78.7	0.0	80.1	75.1	2.0	94.8
Oxford Pets	86.6	0.0	86.8	83.3	3.7	82.7
Caltech101	93.3	0.6	93.8	89.2	3.2	88.6
DTD	49.9	0.0	51.5	49.0	7.3	91.0
EuroSAT	50.8	0.0	41.7	39.1	2.4	89.7
SUN397	65.7	0.0	67.1	64.6	1.1	89.5
UCF101	63.6	0.0	65.3	61.6	4.7	91.4
ImageNet	64.4	0.0	66.3	63.2	0.5	87.1
ImageNet-A	31.1	0.0	32.9	28.9	2.8	89.1
ImageNet-R	68.3	0.1	71.7	68.0	3.7	93.8
ImageNet-Sketch	42.5	0.1	45.1	42.8	0.9	83.1
ImageNetV2	57.2	0.0	59.1	56.0	0.6	85.9

Backbone: ViT-B/32 Attack: PGD-100

Dataset	CLIP		DBD		**D-Pred** (Ours)	
	Clean	Robust	Clean	Robust	Clean	Robust
FGVC Aircraft	20.6	0.0	22.4	65.9	1.9	81.5
Stanford Cars	61.2	0.0	63.7	91.1	0.2	92.4
Flower102	67.0	0.0	68.7	94.0	0.6	95.8
Food101	78.7	0.0	80.1	98.4	2.0	98.6
Oxford Pets	86.6	0.0	86.8	96.7	3.7	88.0
Caltech101	93.3	0.1	93.8	99.2	3.2	96.9
DTD	49.9	0.0	51.5	87.3	7.3	95.3
EuroSAT	50.8	0.0	41.7	92.8	2.4	99.8
SUN397	65.7	0.0	67.1	94.1	1.1	95.2
UCF101	63.6	0.0	65.3	92.4	4.7	95.8
ImageNet	64.4	0.0	66.3	95.4	0.5	94.3
ImageNet-A	31.1	0.0	32.9	95.3	2.8	97.6
ImageNet-R	68.3	0.0	71.7	98.1	3.7	97.7
ImageNet-Sketch	42.5	0.0	45.1	89.7	0.9	90.6
ImageNetV2	57.2	0.0	59.1	94.2	0.6	94.3

C Label-Reliance Controls and Additional Evaluations

C.1 Empirical Checks of Theoretical Quantities

Following Proposition 3.3, we measure M, B, and the observed margin change $\Delta m = m_y(\tilde{\mathbf{x}}) - m_y(\mathbf{x})$ in CLIP fixed-logit-scale units under four label assignments. For the split-sample measurement, two disjoint subsets of low-entropy views, denoted by A and B, produce $G^A = \beta(\rho^A)\min_{c\neq y}(\mathbf{w}_y - \mathbf{w}_c)^\top \bar{\mathbf{d}}^A$ and $\Delta m^B = m_y(\hat{\mathbf{x}} + \beta(\rho^B)\bar{\mathbf{d}}^B) - m_y(\mathbf{x})$. We report G^A, Δm^B, and their correlation, using separate view subsets for predicted gain and observed margin change.

Tables 14, 15, 16, and 17 report the measurements for the quantities in Section 3.3. For PGD-10 and PGD-100, the median residual ratio is below one and the $r < 1$ condition holds for at least 99.5% of samples; their alignment is 91.4 and 94.6, respectively. CW and AA have larger residual ratios and lower alignment. The feature-shift measurements show the same separation: the tangent-space alignment is positive for each attack, while the unprojected alignment is negative. In the margin measurements, GT and pseudo labels give positive median B and positive margin changes, whereas loop and fixed labels give negative median B and near-zero or negative margin changes. The split-sample correlations range from 0.5 to 0.8 across datasets.

Table 14: Measurements of the recovery condition $r = \|\bar{\boldsymbol{\xi}}\|/\|\boldsymbol{\Delta}\|$ on ViT-B/16 fine-grained datasets. Alignment and accuracy are in %.

Attack	Median r	$\Pr[r < 1]$	Alignment	DBD Acc.
CW	1.0	91.1	50.5	56.8
AA	0.8	93.4	61.5	69.1
PGD-10	0.4	99.5	91.4	87.8
PGD-100	0.3	100.0	94.6	91.9

Table 15: Feature-space shift measurements. $P_{\mathbf{x}} = I - \mathbf{x}\mathbf{x}^\top$.

Attack	$\cos(\boldsymbol{\Delta}, \boldsymbol{\mu}(\mathbf{x}) - \mathbf{w}_y)$	$\cos(\boldsymbol{\Delta}, P_{\mathbf{x}}(\boldsymbol{\mu}(\mathbf{x}) - \mathbf{w}_y))$	$\Pr[\mathbf{q}_c^\top \boldsymbol{\Delta} < 0,\ \forall c]$
FGSM	-0.1	0.1	78.3
AA	-0.1	0.1	77.6
PGD-10	-0.2	0.2	91.9
PGD-50	-0.1	0.1	81.7
PGD-100	-0.2	0.2	93.1

Table 16: Margin-certificate measurements on ViT-B/16 fine-grained datasets under PGD-100. $\Pr[B > 0]$ and DBD accuracy are in %.

Label type	Median M	Median B	$\Pr[B > 0]$	Δm	DBD Acc.
GT	11.7	21.5	94.8	7.5	91.9
Pseudo	9.0	15.9	90.9	5.4	67.5
Loop	1.7	-22.0	3.2	-0.1	2.6
Fixed	-34.2	-39.6	0.6	-1.4	0.3

Table 17: Split-sample margins under PGD-100 using disjoint low-entropy view subsets A and B.

Metric	Pets	Cars	Flower102	Food101	Aircraft	Caltech101	SUN397	DTD	EuroSAT	UCF101
G^A	0.4	0.0	0.3	0.1	0.1	0.5	0.3	0.4	0.4	0.5
Δm^B	15.9	14.9	13.0	11.4	14.2	13.3	14.0	10.5	15.2	16.3
$r(G^A, \Delta m^B)$	0.6	0.5	0.7	0.6	0.7	0.7	0.7	0.7	0.8	0.6

C.2 Label-Purity and Conditional Controls

For label-purity control, each attack label is independently replaced by a corrupted label to set the ground-truth label proportion p_{GT}. For conditional controls, we evaluate pseudo-label attacks on

the subsets $\hat{y}_{\text{clean}} = y$ and $\hat{y}_{\text{clean}} \neq y$, where $\hat{y}_{\text{clean}}$ is vanilla zero-shot CLIP's clean prediction, and evaluate GT-guided attacks on the same two subsets.

Table 18 reports the label-purity control, and Table 19 reports conditional accuracy for pseudo-label and GT-guided attacks. As the GT-label proportion decreases from 100% to 0%, DBD accuracy decreases from 92.6 to 0.5. Under pseudo labels, the clean-correct subset has 95.5 average accuracy, whereas the clean-incorrect subset has 1.9. Under GT-guided attacks, the two subset averages remain close (93.0 and 92.3).

Table 18: Label-purity control under PGD-100. Accuracy is in %.

GT-label proportion	100%	80%	60%	40%	20%	0%
Corrupted-label proportion	0%	20%	40%	60%	80%	100%
DBD Acc.	92.6	75.5	47.9	31.8	14.3	0.5

Table 19: Conditional DBD accuracy (%) under PGD-100. Label type specifies the attack label.

Label type	Subset	Pets	Cars	Flower102	Food101	Aircraft	Caltech101	SUN397	DTD	EuroSAT	UCF101	Avg.
Pseudo	$\hat{y}_{\text{clean}} = y$	92.7	93.0	97.2	99.6	86.9	98.7	97.1	97.1	95.3	97.3	95.5
	$\hat{y}_{\text{clean}} \neq y$	1.1	6.1	0.3	0.1	2.7	5.6	0.6	0.4	1.3	0.6	1.9
GT	$\hat{y}_{\text{clean}} = y$	93.2	97.1	89.3	97.2	87.9	88.6	83.8	99.3	97.4	96.3	93.0
	$\hat{y}_{\text{clean}} \neq y$	93.0	96.3	84.7	99.3	92.1	93.9	84.6	95.2	88.0	96.3	92.3

C.3 Perturbation-Budget Controls

We evaluate PGD-100 with $\varepsilon \in \{4/255, 6/255, 8/255, 12/255\}$ under GT, pseudo, loop, and fixed label assignments. For GT and pseudo labels, both DBD and D-Pred retain high accuracy across all four budgets and increase with ε in these attack settings. Loop and fixed labels remain near zero for both methods at every budget, so changing the perturbation budget does not recover the GT-guided pattern for these label assignments.

Table 20: Accuracy (%) under PGD-100 at four perturbation budgets.

Label type	Method	$\varepsilon = 4/255$	$\varepsilon = 6/255$	$\varepsilon = 8/255$	$\varepsilon = 12/255$
GT	DBD	92.8	95.8	96.1	99.3
	D-Pred	91.2	95.3	96.7	98.9
Pseudo	DBD	65.3	67.1	67.2	68.1
	D-Pred	68.5	69.2	70.7	70.5
Loop	DBD	0.8	0.9	1.4	1.8
	D-Pred	0.6	1.0	1.2	1.4
Fixed	DBD	2.2	1.9	1.7	1.6
	D-Pred	1.5	1.4	1.6	1.4

C.4 Transfer to Other Test-Time Defenses

We append DBD to R-TPT, TAPT, and TTC and evaluate the resulting pipelines under GT-guided PGD-100. Table 21 reports the per-dataset and average accuracies. Appending DBD raises their average accuracies from 56.9, 49.5, and 68.9 to 89.2, 91.4, and 91.1, respectively. The gains occur across all ten datasets; for example, R-TPT on Aircraft increases from 14.8 to 92.1, and TAPT on Caltech101 increases from 13.7 to 95.0.

Table 21: Accuracy (%) of test-time defenses with and without DBD under GT-guided PGD-100.

Method	Pets	Cars	Flower102	Food101	Aircraft	Caltech101	SUN397	DTD	EuroSAT	UCF101	Avg.
R-TPT	70.5	66.9	58.6	62.3	14.8	83.7	58.1	47.5	43.7	63.2	56.9
+ DBD	91.3	89.1	86.7	97.2	92.1	85.4	81.0	88.9	94.1	85.9	89.2
TAPT	72.4	52.7	44.3	50.3	61.2	13.7	72.1	32.6	40.3	55.8	49.5
+ DBD	86.7	90.2	95.3	89.2	85.3	95.0	97.4	88.1	96.7	90.3	91.4
TTC	60.9	57.7	78.0	76.0	49.0	95.5	78.9	68.4	51.0	73.6	68.9
+ DBD	90.8	88.9	96.7	87.3	92.0	95.6	90.9	89.5	94.5	85.0	91.1

We also evaluate R-TPT, TTC, DOC, TAPT, and AOM with and without D-Pred under GT, loop, and fixed label assignments, as reported in Table 22. For D-Pred, the GT rows increase all five pipelines, whereas the loop and fixed rows have smaller and non-uniform changes across pipelines.

Table 22: Mean test-time defense accuracy (%) with and without D-Pred for three label settings.

Label setting	R-TPT	TTC	DOC	TAPT	AOM
GT	56.9	68.9	55.6	49.5	58.4
+ D-Pred	81.9	86.5	76.8	80.6	83.2
Loop	68.2	57.5	53.7	72.1	64.9
+ D-Pred	70.4	72.6	57.8	76.1	66.2
Fixed	66.8	65.3	56.6	77.3	63.3
+ D-Pred	71.0	67.5	74.6	75.9	62.9

C.5 Complete-Pipeline Adaptive Evaluation

For complete-pipeline attacks, PGD maximizes the final defended classification loss through the transformations, entropy selection, direction estimation, and feature reconstruction. We use EOT and BPDA (Athalye et al., 2018) with PGD-100, $\varepsilon = 4/255$, step size $1/255$, four EOT samples, and three restarts. Table 23 reports the per-dataset and average accuracies.

Relative to the non-adaptive attack on vanilla CLIP, the adaptive attack reduces the average by 3.7 points for DBD and 2.7 points for D-Pred. The per-dataset values are lower except for D-Pred on DTD, which changes from 91.0 to 91.1; Aircraft has the lowest adaptive accuracy for both pipelines (73.7 for DBD and 73.3 for D-Pred).

Table 23: Accuracy (%) of DBD and D-Pred under PGD-100. Columns indicate the attack target: vanilla CLIP (non-adaptive) or the full defense pipeline (adaptive). Adaptive attacks use EOT and BPDA with four EOT samples and three restarts.

	DBD		D-Pred	
Dataset	CLIP	Full pipeline	CLIP	Full pipeline
Pets	96.7	91.3	93.5	90.7
Cars	93.5	90.2	92.7	89.9
Flower102	98.7	93.4	95.6	92.7
Food101	94.3	91.2	93.8	91.7
Aircraft	76.0	73.7	75.2	73.3
Caltech101	97.1	93.9	95.2	92.2
SUN397	93.1	87.6	92.4	86.1
DTD	92.1	90.4	91.0	91.1
EuroSAT	96.3	94.3	95.1	92.6
UCF101	95.9	90.9	93.6	90.8
Avg.	**93.4**	**89.7**	**91.8**	**89.1**