

USENIX

THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

Low-Cost Hard-Label Adversarial Attack with Theoretical Foundations

Jun Liu, *University of Macau; National Institute of Informatics;* Leo Yu Zhang, *Griffith University;* Fengpeng Li, *University of Macau;* Isao Echizen, *University of Tokyo; National Institute of Informatics;* Jiantao Zhou, *University of Macau*

https://www.usenix.org/conference/usenixsecurity26/presentation/liu-jun

This paper is included in the Proceedings of the 35th USENIX Security Symposium.

August 12–14, 2026 • Baltimore, MD, USA

ISBN 978-1-939133-58-8

Open access to the Proceedings of the 35th USENIX Security Symposium is sponsored by



جامعة الملك عبدالله للعلوم والتقنية

King Abdullah University of Science and Technology

ARTIFACT EVALUATED
USENIX
AVAILABLE

Low-Cost Hard-Label Adversarial Attack with Theoretical Foundations

Jun Liu[†‡], Leo Yu Zhang[§], Fengpeng Li[†], Isao Echizen[¶‡], Jiantao Zhou[†*]

[†]*University of Macau,* [‡]*National Institute of Informatics,* [§]*Griffith University,* [¶]*University of Tokyo*
[*]*Corresponding author*

Abstract

Hard-label black-box attacks, relying solely on top-1 predictions, represent one of the most challenging yet practically threat models. Despite recent progress, existing approaches face two key limitations: (1) they overlook the critical role of initialization, focusing primarily on optimization strategies; and (2) they rely heavily on empirical heuristics without theoretical guarantees. To bridge this gap, we establish a unified theoretical framework showing that existing sign-flipping hard-label attacks can be understood as approximating the true gradient sign. Guided by this principled analysis, we propose a novel attack framework featuring a zero-query initialization strategy and a Pattern-Driven Optimization (PDO) algorithm. We provide theoretical guarantees that our initialization yields higher cosine similarity to the true gradient sign than random baselines, and our PDO module achieves significantly lower query complexity than baseline search methods. Extensive experiments across CIFAR-10, ImageNet, and ObjectNet—covering standard and adversarially trained models, commercial APIs, and CLIP models—demonstrate that our method consistently outperforms SOTA hard-label attacks in both success rate and efficiency, particularly under low query budgets. Furthermore, our method demonstrates robust generalization across corrupted data (ImageNet-C), biomedical images (PathMNIST), and dense prediction tasks such as segmentation. Notably, it bypasses the stateful defense Blacklight, achieving a 0% detection rate.

1 Introduction

Deep Neural Networks (DNNs) have been widely applied in real-world applications such as image classification [4, 24], speech recognition [64], face recognition [18, 34], and etc. However, previous researches [22,37,38,58] reveal that DNNs can fail catastrophically when processing Adversarial Examples (AEs)—carefully crafted inputs that deceive DNNs while appearing benign to human observers. In image classification tasks, adversarial attacks for generating AEs can be categorized based on the adversary's knowledge of the victim model.

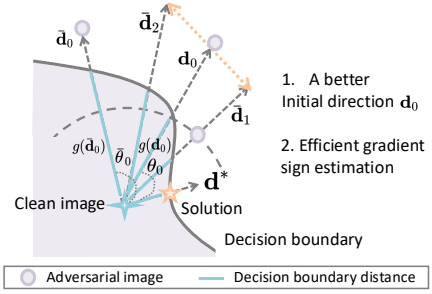


Figure 1: Comparison of our approach with traditional methods. Our method achieves a better initialization $\mathbf{d}_0$ with smaller angular deviation θ_0 and boundary distance $g(\mathbf{d_0})$ compared to the traditional initialization $\bar{\mathbf{d}}_0$. Furthermore, it employs a more query-efficient strategy to approximate the ideal descent direction $\mathbf{d}^*$.

Broadly, they fall into three classes: white-box [15, 19, 63], gray-box [26], and black-box attacks [2,7,56]. In *white-box* attacks, the adversary has complete access to the victim model's architecture and parameters, while *gray-box* attacks assume partial access. In contrast, *black-box* attacks assume no knowledge of either the architecture or internal parameters of the victim model and can be further classified by the type of output they leverage: *soft-label* (namely score-based) attacks [1, 25] exploiting classification confidence scores (e.g., softmax outputs) and *hard-label* (a.k.a. decision-based) attacks [8, 53, 55] relying solely on the top-1 predicted labels. Among these attacks, hard-label attacks represent the most challenging and practical scenario [8, 11, 12, 55, 56]. Real-world applications—such as industry AI systems (e.g., in medical imaging [16]) and commercial APIs (e.g., Core ML [3])—intentionally suppress output confidences due to data privacy regulations and mobile resource constraints, thereby necessitating decision-only attack strategies and making them crucial for evaluating and improving the robustness of DNNs. Hence, this paper primarily focuses on hard-label attacks.

Traditional hard-label attack workflow. Current hard-label attacks are broadly divided into surrogate-assisted and surrogate-free methods. Surrogate-assisted methods (or transfer-query attacks) leverage gradient priors from local models to guide the search [49]. Although efficient, their performance degrades significantly under architectural mismatch (e.g., Convolution Neural Networks (CNNs) vs. Transformers) or domain shifts, and can be mitigated by ensemble training [51]. In contrast, surrogate-free (query-based) methods operate independently of external priors, relying solely on the victim model's hard-label feedback to estimate gradients or refine perturbations. The workflow of these query-based methods typically involves two stages:

Stage 1: Perturbation direction initialization. This stage aims to generate an initial search direction. Traditional methods randomly sample noise or another image until an AE is found. The vector from the clean image to this AE defines the initial direction. Then, a binary search is performed along this vector to locate a *starting point* close to the decision boundary.

Stage 2: Direction refinement and perturbation reduction. Starting from the boundary point, this stage iteratively refines the initial direction and reduces the perturbation magnitude based on hard-label feedback. The objective is to identify an AE that satisfies the specified perturbation constraints. If no valid example is found within the allocated query budget, the attack is deemed unsuccessful.

Observations and insights. 1) Better initialization. We observe that most existing black-box attacks devote substantial effort to Stage 2 (the search phase), while largely overlooking the importance of Stage 1 (initialization). In practice, they typically rely on random strategies, such as Gaussian or Uniform noise [8, 13, 43, 55], random images [42, 55], or all **1** perturbations [9, 56]. These choices ignore prior knowledge about model vulnerabilities, often producing initial directions that are far from the decision boundary and difficult to refine under limited query budgets (e.g., < 100 queries). This motivates us to design a prior-guided initialization that yields a superior starting direction $\mathbf{d}_0$. As illustrated in Fig. 1, our goal is to obtain an initialization that is closer to the ideal descent direction $\mathbf{d}^*$, both in angular deviation and in decision-boundary distance, compared to conventional random starts. **2) More efficient gradient sign estimation.** The next challenge is to efficiently approach the ideal descent direction $\mathbf{d}^*$. Inspired by the score-based method SignHunter [1], we identify $\mathbf{d}^*$ with the true gradient sign utilized in FGSM [30], a pragmatic approximation given that exact gradient retrieval is intractable in the black-box setting. Specifically, SignHunter demonstrates that, with sufficient queries, it can approximate the gradient sign used in FGSM. Similarly, we observe that RayS [9] can asymptotically achieve FGSM performance by iteratively refining sign estimates through sign-flipping. However, in practice, the efficiency of such methods deteriorates rapidly as the dimensionality grows. Moreover, when refining our structured initial directions, RayS-style rigid equal partitioning ignores the underlying spatial pattern, frequently breaking coherent regions into fragmented pieces. This leads to *gain cancellation*: when a partition contains conflicting gradient signs, their contributions offset one another, thereby weakening the effective descent signal and slowing the search process. These observations suggest that efficient gradient sign approximation requires not only reducing the search dimensionality, but also preserving spatial coherence throughout the search process.

Our work. From a *greedy optimization* perspective, we propose an efficient black-box attack that jointly improves initialization and search. First, we design a superior initialization strategy via Block-wise Discrete Cosine Transform (BDCT) analysis [54]. By characterizing model sensitivity across frequency bands, we derive a data-driven sampling distribution from the clean image's frequency statistics and generate an initial direction via inverse BDCT. Both theoretical analysis and empirical results show that this initialization is significantly closer to the decision boundary and exhibits higher cosine similarity with the true gradient sign than random baselines (Sec. 4.3). Building upon this structured initialization, we further introduce a *Pattern-Driven Optimization* (PDO) module. PDO explicitly preserves the spatial coherence of initial directions throughout the search process, thereby reducing the effective search dimensionality and mitigating gain cancellation. As a result, it approximates the true gradient sign with substantially fewer queries than the equal partitioning searches used in state-of-the-art (SOTA) baselines (Sec. 4.4).

We summarize our main contributions as follows:

- We provide the first theoretical analysis showing that RayS-like sign-flipping hard-label black-box attacks can asymptotically match white-box FGSM. Our analysis identifies search dimensionality and cumulative sign-estimation errors as the primary bottlenecks limiting practical efficiency.
- We propose a zero-query initialization strategy grounded in image frequency statistics. Unlike blind random sampling, this method leverages prior knowledge to locate a starting point with a smaller decision boundary distance and better alignment with the true gradient sign.
- We design the PDO method, which reduces the search space and mitigates the gain cancellation effect inherent in baseline equal-partitioning search. We provide theoretical guarantees on its improved query complexity and experimentally demonstrate its faster convergence and higher cosine similarity with the true gradient sign.
- Extensive experiments demonstrate that our proposed DPAttack significantly outperforms existing algorithms in terms of ASR and query efficiency. It proves highly effective against traditional DNNs, foundation models (e.g., CLIP), and real-world APIs, while successfully bypassing SOTA defenses like Blacklight.

2 Related Work

2.1 Problem Formulation & Background

Given a victim model $\mathcal{F}: \mathbb{R}^d \to \{1, \cdots, M\}$ (where M denotes the number of classes), a clean image $\mathbf{x} \in \mathbb{R}^d$, and its ground-truth (GT) label y, an untargeted adversarial attack aims to generate an adversarial example (AE) $\mathbf{x}^* = \mathbf{x} + \Delta$ such that $\mathcal{F}(\mathbf{x}^*) \neq y$. The objective is to maximize the classification loss $\mathcal{L}(\mathbf{x}+\Delta, y)$ while constraining the perturbation magnitude within an allowed ℓ_p-ball (typically $p \in \{2, \infty\}$) of radius ε. Formally:

$$\max_{\Delta} \mathcal{L}(\mathbf{x}+\Delta, y), \quad \text{s.t.} \quad \|\Delta\|_p \leq \varepsilon. \tag{1}$$

In white-box settings, attackers optimize Eq. (1) using gradients, e.g., the Fast Gradient Sign Method (FGSM) [30]: $\mathbf{x}' = \mathbf{x} + \varepsilon \cdot \text{sgn}(\nabla \mathcal{L}(\mathbf{x}, y))$. However, in the hard-label black-box setting, the attacker only accesses the top-1 predicted label, rendering direct gradient access impossible.

2.2 Hard-Label Black-Box Attacks

Early research, such as the Boundary Attack [7], relied on heuristic random walks along the decision boundary. While pioneering, it suffered from extremely low query efficiency due to the blindness of its search strategy. To improve convergence, subsequent works incorporated gradient or normal vector estimation to guide the attack. Canonical methods like HSJA [8] and Bounce [55] utilize Monte Carlo sampling to approximate the decision boundary normal, enabling more directed perturbation updates. Building on this, TtBA [57] weights the perturbation direction with its estimated normal vector. It utilizes a robust "two-third" update rule and curvature-based detection to escape local optima, achieving superior efficiency in normal-vector-based attacks.

However, these gradient-estimation-based methods typically require substantial query budgets to approximate normal vectors. Distinct from this direction, a parallel line of optimization-based approaches bypasses explicit normal estimation. Opt [13] pioneered this by formulating the attack as a zero-order optimization task. In line with this approach, SignOpt [12] further improved query efficiency by reconstructing the gradient direction solely from the aggregated signs of directional derivatives along random projections, rather than estimating magnitude. Specifically, SignOpt reformulates the attack objective in Eq. (1) as finding a unit vector $\mathbf{v} \in \mathbb{S}^{d-1}$ (where $\mathbb{S}^{d-1} := \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_2 = 1\}$) to minimize:

$$g(\mathbf{v}) = \min\{r > 0 \mid \mathcal{F}(\mathbf{x} + r \cdot \mathbf{v}) \neq y\}. \tag{2}$$

Here, $g(\cdot)$, defined on $\mathbb{S}^{d-1}$, represents the shortest distance r from the clean image $\mathbf{x}$ to the decision boundary along direction $\mathbf{v}$. SignOpt solves Eq. (2) by estimating the gradient $\nabla_{\mathbf{v}} g(\mathbf{v})$ and updating $\mathbf{v}$ using standard Stochastic Gradient Descent (SGD) as follows:[1]

1. Sample multiple directions $\{\mathbf{n}_i\}_{i=1}^N$ from a Gaussian distribution, typically setting $N = 200$ to mitigate noise.
2. Obtain the sign of the directional derivative via finite differences:[2] $\text{sgn}(\nabla_{\mathbf{n}_i} g(\mathbf{v})) = \text{sgn}([g(\mathbf{v} + \tau \mathbf{n}_i) - g(\mathbf{v})]/\tau)$, where τ is a small constant.
3. Approximate $\nabla_{\mathbf{v}} g(\mathbf{v}) \approx \sum_{i=1}^N \text{sgn}(\nabla_{\mathbf{n}_i} g(\mathbf{v})) \cdot \mathbf{n}_i$ and update $\mathbf{v}$ using SGD with a learning rate η.

Despite these advancements, continuous gradient-recovery approaches still struggle with high-dimensional search spaces, particularly under ℓ_∞ constraints. This limitation motivated the development of following SOTA discrete search strategies.

2.3 Discrete Hard-Label Attacks

Unlike the aforementioned methods that attempt to approximate continuous gradients, a distinct category of attacks, represented by RayS [9] and ADBA [56], abandons explicit gradient estimation. Instead, these methods directly search for the optimal perturbation direction within a discrete space (i.e., $\{-1, 1\}^d$), leveraging the empirical observation that ℓ_∞-bounded adversarial perturbation often lies on the vertices of the ℓ_∞ ball [44] to accelerate the search process. As our proposed framework follows and improves upon this discrete search paradigm, we review these two baselines in detail.

RayS [9]. RayS converts continuous optimization into a discrete search problem over $\mathcal{H} \equiv \{-1, +1\}^d$. To this end, its optimization objective is:

$$\min_{\mathbf{d} \in \mathcal{H}} g(\mathbf{d}) \quad \text{where } \mathcal{H} \equiv \{-1, +1\}^d. \tag{3}$$

Naive RayS (NRayS) heuristically optimizes Eq. (3) by flipping the sign of each dimension in the search space $\mathcal{H}$, while maintaining and updating the minimum perturbation magnitude $g(\mathbf{d}_{\text{best}})$ and its corresponding direction $\mathbf{d}_{\text{best}}$. It initializes the direction as an all-one vector, i.e. $\mathbf{d}_{\text{best}} = \mathbf{d}_0 = \mathbf{1}$, and identifies $g(\mathbf{d}_0)$ via binary search.

To circumvent the inefficiency of dimension-wise flipping in NRayS, the authors proposed Hierarchical RayS (HRayS). HRayS iteratively refines $\mathbf{d}_{\text{best}}$ through a hierarchical dyadic structure, effectively exploiting spatial correlations in image gradients [32]. Specifically, HRayS constructs a standard dyadic search tree T_{dyad} via the following three steps:

Step 1. Index partitioning. At each hierarchy level $l \in \{1, 2, \cdots, \lceil \log_2 d \rceil\}$, the index set $\Gamma = \{1, 2, \cdots, d\}$ of vector $\mathbf{d}_{\text{best}}$ is sequentially partitioned into 2^l disjoint subsets $\{B_1^{(l)}, B_2^{(l)}, \cdots, B_{2^l}^{(l)}\}$ of equal length $d/2^l$.

Step 2. Sign flipping and direction update. For each subset $B_k^{(l)}$ ($k \in [1, 2^l]$), the signs of the entries in $\mathbf{d}_{\text{best}}$ corresponding to this subset are flipped, yielding a perturbed direction $\mathbf{d}_k^{(l)}$.

[1] The minimization over r is performed using a binary search.

[2] Note that calculating this directional derivative only requires access to the hard-label oracle $\mathcal{F}(\mathbf{x} + r\mathbf{v}) \neq y$.

The current best direction $\mathbf{d}_{\text{best}}$ is updated *only if* the flipped direction yields a smaller boundary distance. That is

$$\mathbf{d}_{\text{best}} = \begin{cases} \mathbf{d}_k^{(l)} & \text{if } g(\mathbf{d}_k^{(l)}) < g(\mathbf{d}_{\text{best}}) \\ \mathbf{d}_{\text{best}} & \text{Otherwise.} \end{cases} \tag{4}$$

After traversing all 2^l subsets, $B_{ks}^{(l)}$ is defined as the union of subsets chosen for updating $\mathbf{d}_{\text{best}}$.

Step 3. Refinement. Steps 1 and 2 are repeated by increasing l by 1, progressively refining the direction at finer granularities, until $l = \lceil \log_2 d \rceil$, the query limit is reached, or $g(\mathbf{d}_{\text{best}})$ satisfies the perturbation budget.

ADBA [56]. As the current SOTA for untargeted ℓ_∞ attacks, ADBA follows the hierarchical structure of RayS but identifies a critical inefficiency: RayS relies on binary search to calculate precise boundary distances $g(\cdot)$ for every candidate, which is query-expensive. ADBA improves this by comparing *pairs* of search directions.

Step 1. Index partitioning: Identical to Step 1 in HRayS.

Step 2. Sign flipping and direction update. For each pair of adjacent subsets $(B_{2k-1}^{(l)}, B_{2k}^{(l)})$ where $k \in [1, 2^{l-1}]$, the signs of entries in $\mathbf{d}_{\text{best}}$ (initialized as $\mathbf{d}_0 = \mathbf{1}$) corresponding to these two groups are flipped respectively, yielding two candidate directions $\bar{\mathbf{d}}_1$ and $\bar{\mathbf{d}}_2$. A comparative function $\Lambda(\cdot,\cdot)$ determines which direction is closer to the boundary only when both candidates are adversarial:

$$\mathbf{d}_{\text{best}} = \begin{cases} \mathbf{d}_{\text{best}} & \text{if } \mathcal{F}(\mathbf{x}+r\bar{\mathbf{d}}_1) = y \;\&\; \mathcal{F}(\mathbf{x}+r\bar{\mathbf{d}}_2) = y \\ \bar{\mathbf{d}}_1 & \text{if } \mathcal{F}(\mathbf{x}+r\bar{\mathbf{d}}_1) \neq y \;\&\; \mathcal{F}(\mathbf{x}+r\bar{\mathbf{d}}_2) = y \\ \bar{\mathbf{d}}_2 & \text{if } \mathcal{F}(\mathbf{x}+r\bar{\mathbf{d}}_1) = y \;\&\; \mathcal{F}(\mathbf{x}+r\bar{\mathbf{d}}_2) \neq y \\ \Lambda(\bar{\mathbf{d}}_1, \bar{\mathbf{d}}_2) & \text{Otherwise,} \end{cases} \tag{5}$$

where r denotes the current minimal perturbation distance (initialized to 1 for $[0,1]$-valued images), and $\Lambda(\cdot,\cdot)$ returns the direction with the smaller decision boundary distance. The function Λ iteratively checks whether perturbing $\mathbf{x}$ by a gradually reduced r along both $\bar{\mathbf{d}}_1$ and $\bar{\mathbf{d}}_2$ maintains adversarial status. This process stops at r^* if a direction fails to produce an AE or a step limit (e.g., 8) is reached. After determining $\mathbf{d}_{\text{best}}$ and updating r to r^*, the algorithm returns to Step 1.

Step 3. Refinement. Identical to Step 3 in HRayS.

Limitation: Although ADBA substantially improves query efficiency over RayS, it relies on a naive all-ones initialization ($\mathbf{d}_0 = \mathbf{1}$), which limits its starting efficiency. More importantly, ADBA remains a purely empirical method that lacks theoretical guarantees regarding its convergence or gradient approximation accuracy. In contrast, our work establishes a theoretically grounded framework that combines a frequency-guided initialization with a structured, pattern-aware optimization strategy, yielding both stronger guarantees and higher practical efficiency.

3 Theoretical Foundations: From Discrete Search to Gradient Sign Estimation

To address the lack of theoretical guarantees in RayS-like hard-label attacks, we draw inspiration from their soft-label counterpart SignHunter [1]. While SignHunter is proven to approximate FGSM, its proof relies on score-based derivative estimation, which is unavailable in hard-label settings. To bridge this gap, we establish a novel theoretical framework demonstrating that RayS-based algorithms (specifically NRayS and HRayS) asymptotically approximate FGSM. As ADBA is a derivative of these core algorithms, our analysis focuses on the fundamental RayS mechanism. We first derive a probabilistic lower bound for the alignment between a random binary vector and the true gradient sign. This provides a key building block for understanding why discrete sign-search methods can recover meaningful gradient information in high dimensions.

Theorem 1 (Approximation Lower Bound for NRayS)**.** *Let $\mathbf{u} \in \mathbb{S}^{d-1}$ be an arbitrary unit vector. Let $\{\mathbf{d}_j\}_{j=1}^m$ be m vectors sampled independently and uniformly from the binary hypercube $\mathcal{H} \equiv \{-1,+1\}^d$, and let $\hat{\mathbf{d}}_j = \mathbf{d}_j/\sqrt{d}$ be their normalized counterparts. For any precision $\zeta \in (0,1)$ and failure probability $\delta \in (0,1)$, there exists a constant $C \geq 1$ such that if the sample size satisfies $m \geq -(\ln \delta) e^{Cd\zeta^2}$, then with probability at least $1-\delta$, there exists at least one index $j \in \{1,\ldots,m\}$ satisfying:*

$$\mathbf{u}^\top \hat{\mathbf{d}}_j \geq \zeta. \tag{6}$$

Proof. Due to space constraints, the proofs are provided in Appendix (App.) A of our arXiv version [41]; the same applies to subsequent proofs. By instantiating Theorem 1 with the normalized gradient $\mathbf{u} = \nabla\mathcal{L}/\|\nabla\mathcal{L}\|_2$, we conclude that *NRayS, given a sufficient query budget m, guarantees finding a direction $\mathbf{d} \in \mathcal{H}$ that aligns with the true gradient $\nabla\mathcal{L}$ with high probability, thereby approximating the FGSM update.* While the theoretical bound implies high query complexity (e.g., assuming $C = 1$, ensuring $\zeta \geq 0.1$ with 50% probability on $3 \times 32 \times 32$ sized CIFAR-10 [36] images requires $m \approx 1.5^{13}$ queries), empirical findings from SignHunter suggest that strict gradient alignment is unnecessary. In practice, partial alignment—recovering the signs of the top $20-30\%$ of gradients—suffices to achieve attack success rates exceeding 70% on datasets like CIFAR-10 and ImageNet [17].

Theorem 1 establishes the theoretical feasibility of using discrete search to approach the true gradient sign. However, purely random search remains inefficient in high dimensions. Extending this insight, Theorem 2 demonstrates how HRayS leverages structured estimation to bypass this bottleneck. It shows that HRayS constructs a direction aligned with the gradient sign, where the approximation quality is governed by the subset selection error rather than blind sampling.

Definition 1 (Subset-wise Directional Derivative Influence)**.** *For a selected index subset $B_{ks}^{(l)}$ at round l of HRayS (following the definition in Step 2 of HRayS; see Sec. 2.3), we define its influence as the cumulative magnitude of the gradient components within that subset:*

$$I(B_{ks}^{(l)}) := \sum_{i \in B_{ks}^{(l)}} \left| \frac{\partial \mathcal{L}(\tilde{\mathbf{x}}, y)}{\partial \tilde{\mathbf{x}}[i]} \Big|_{\tilde{\mathbf{x}}=\mathbf{x}+g(\mathbf{d}_{best})\mathbf{d}_{best}} \right|. \tag{7}$$

This metric quantifies the collective contribution of dimensions in $B_{ks}^{(l)}$ to increasing the loss. Based on this influence measure, we establish the following subset selection guarantee for HRayS.

Lemma 1. *Under standard regularity conditions, the subset $B_{ks}^{(l)}$ selected by HRayS at round l satisfies:*

$$I(B_{ks}^{(l)}) \geq \max_{B_k^{(l)}} I(B_k^{(l)}) - \eta_l, \tag{8}$$

where η_l represents the selection error bound at round l.

Proof. Provided in App. B of our arXiv version [41]. This result guarantees that each HRayS round identifies a high-impact subset, enabling the cumulative progress analysis in the following theorem.

Theorem 2 (Gradient Approximation Guarantee for HRayS)**.** *Assume that the gradient magnitude is uniformly bounded, i.e., $\|\nabla \mathcal{L}(\mathbf{x}, y)\|_\infty \leq G_{\max}$. Let η_l denote the maximum deviation between the directional derivative influence of the subset $B_{ks}^{(l)}$ selected by HRayS at round l and that of the optimal subset $\arg\max_{B_k^{(l)}} I(B_k^{(l)})$. Let $\hat{\mathbf{d}}^*$ be the final direction produced by HRayS. Then, the alignment between $\hat{\mathbf{d}}^*$ and the true gradient sign satisfies*

$$\operatorname{sgn}(\nabla \mathcal{L}(\mathbf{x}, y))^\top \hat{\mathbf{d}}^* \geq \frac{\|\nabla \mathcal{L}(\mathbf{x}, y)\|_1}{G_{\max}} \cdot (1 - \varpi), \tag{9}$$

where $\varpi = \sum_{l=1}^{\lceil \log_2 d \rceil} \eta_l \|\nabla \mathcal{L}(\mathbf{x}, y)\|_1^{-1}$ denotes the cumulative relative error induced by suboptimal subset selections across all hierarchical rounds.

Proof. Provided in App. B of our arXiv version [41]. Theorem 2 reveals that although HRayS does not explicitly estimate the gradient, its hierarchical decision mechanism effectively mimics the selection of dimensions with the steepest ascent. By progressively refining the coordinate set, HRayS constructs a direction $\hat{\mathbf{d}}^*$ that closely aligns with $\operatorname{sgn}(\nabla \mathcal{L}(\mathbf{x}, y))$, provided the selection error ϖ remains small.

Implications for Our Approach. Theorems 1 and 2 suggest that the efficiency of hard-label attacks hinges on two factors: (1) the search space dimensionality (governing the bound in Theorem 1), and (2) the cumulative selection error (quantified by ϖ in Theorem 2). Motivated by these theoretical insights, we propose DPAttack, which bridges these gaps through two principled design choices:

Frequency-Prior Initialization (Sec. 4.3): To mitigate the high dimensionality penalty, we utilize frequency-domain priors to design a zero-query initialization. Unlike random noise, this direction exhibits local spatial coherence (i.e., blocks of consistent signs), effectively reducing the search complexity from the full pixel space to a structured subspace. Theorem 3 further guarantees that this method yields a positive expected alignment with the true gradient sign.

Pattern-Driven Optimization (Sec. 4.4): To address the cumulative error ϖ (Theorem 2), we identify that the fixed partitioning in RayS and ADBA often disrupts the spatial coherence established by our initialization, leading to conflicting updates (gain cancellation). We thus employ a tailored pattern-driven search that groups pixels by their initial sign coherence. By respecting initialization's intrinsic structures, this strategy is proven to dominate baselines in sign alignment (Theorem 4) and reduce query complexity (Theorem 5).

4 Our Proposed Attack

4.1 Threat Model

We align our threat model with standard hard-label black-box attack settings [9], defined by the following aspects:

Adversary's Goal. The adversary aims to generate an untargeted adversarial example $\mathbf{x}^*$ that misclassifies the victim model (i.e., $\mathcal{F}(\mathbf{x}^*) \neq y$) while maintaining the perturbation within an imperceptible budget (e.g., satisfying ℓ_∞ or ℓ_2-norm constraints). Formally, we adopt the same optimization objective as RayS (defined in Eq. (3)), minimizing the distance to the decision boundary within the discrete search space.

Adversary's Knowledge and Capabilities. We assume a hard-label black-box scenario. The adversary can only interact with the victim model $\mathcal{F}$ via an oracle (e.g., an online API) to obtain the top-1 predicted label for a given input. Access to the model's architecture, parameters, gradients, or continuous output probabilities (soft labels) is strictly prohibited.

4.2 The Overall Attack Pipeline of DPAttack

Guided by the theoretical insights established in Sec. 3, we propose DPAttack, a coarse-to-fine framework designed to efficiently approximate the true gradient sign in FGSM. As illustrated in Fig. 2 and Algorithm 1, the pipeline consists of two integrated stages: (1) The Dynamic Decision-Making (DDM) Module (Sec. 4.3), which leverages frequency priors to generate a structured and informative initialization; and (2) The Pattern-Driven Optimization (PDO) Module (Sec. 4.4), which refines this direction by exploiting spatial coherence to accelerate convergence. We detail these components below.

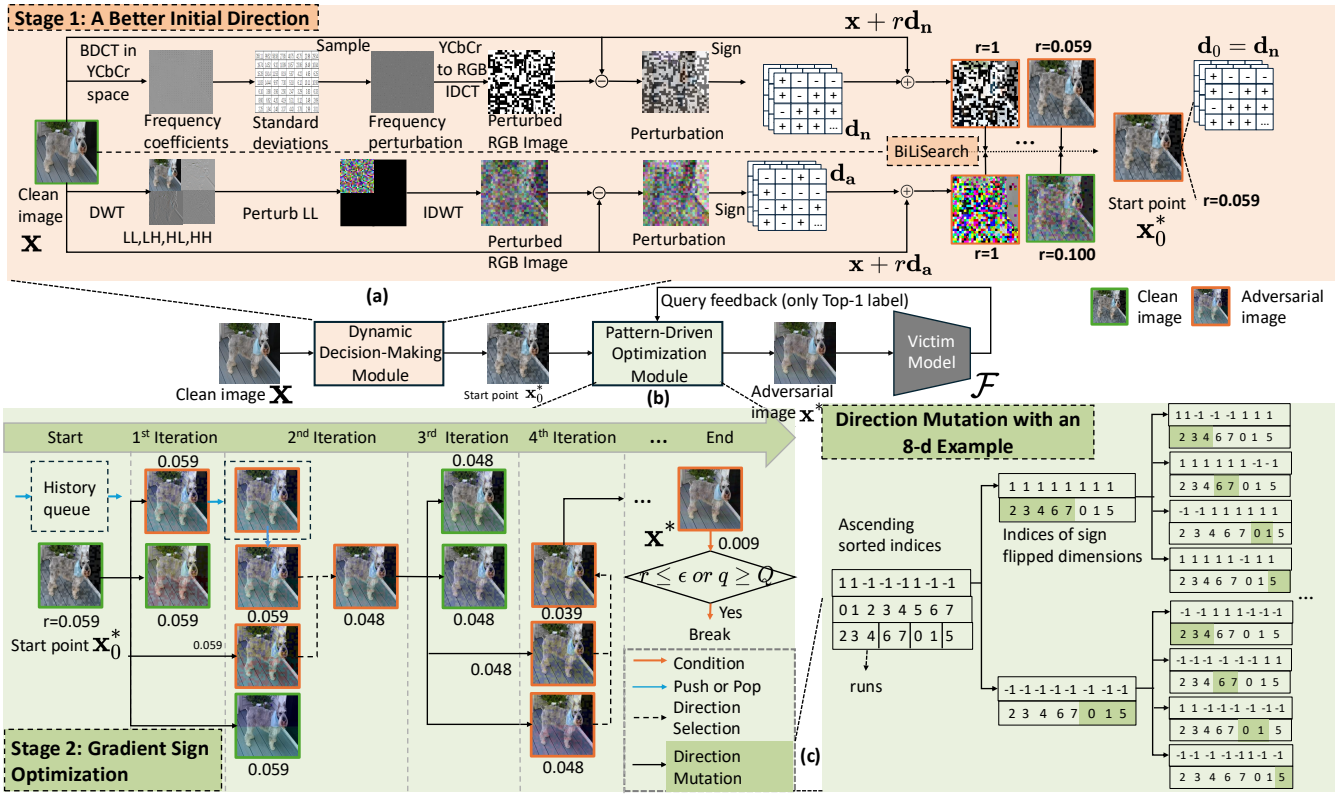


Figure 2: Framework of the proposed hard-label attack method DPAttack. (a) Stage 1: The Dynamic Decision-Making (DDM) module generates a superior initial perturbation direction $\mathbf{d}_0$. (b) Stage 2: The Pattern-Driven Optimization (PDO) module iteratively refines the direction and the perturbation magnitude r, yielding the final adversarial example $\mathbf{x}^* = \text{clip}(\mathbf{x} + r\hat{\mathbf{d}}^*, [0,1])$.

4.3 Better Initialization via DDM

Our initialization approach relies on the premise that profiling model vulnerabilities provides effective prior knowledge for constructing potent initial perturbations. While previous studies [10] have used Shapley values to identify frequency sensitivity, this metric is derived by masking frequency bands rather than perturbing them, making it ill-suited for capturing sensitivity to additive noise. To address this, we introduce a novel analysis method and propose the DDM module.

4.3.1 Analysis: Block-DCT Frequency Sensitivity (BFS)

To accurately characterize model vulnerability to frequency-based noise, we propose BFS. While drawing conceptual inspiration from the frequency sensitivity analysis in [62], our method significantly advances this framework by introducing two critical modifications tailored for adversarial efficiency: 1) We employ the variation in CE loss rather than classification error rate to provide a fine-grained sensitivity metric; 2) We apply perturbations in the Block Discrete Cosine Transform (BDCT) domain [54] instead of the Fast Fourier Transform domain [23], as DNNs are typically more vulnerable to the block-wise artifacts introduced by BDCT.

BFS Procedure. Given an RGB image $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$, we convert it to YCbCr space and divide it into non-overlapping blocks of size $w \times w$ (e.g., $w = 8$). Performing DCT on these blocks yields the frequency matrix $\mathbf{I} \in \mathbb{R}^{C\times Z\times w\times w}$, where $Z = (H/w)\times(W/w)$ is the total number of blocks. To measure sensitivity, we define a Gaussian noise distribution $\mathcal{N}(0,\sigma^2_{\max})$, where $\sigma_{\max}$ is calibrated such that the perturbation exhausts the pixel-domain budget ε. For a specific frequency coordinate (c,i,j), we sample a noise vector $\hat{\mathbf{w}}_{c,i,j} \in \mathbb{R}^Z$ from this distribution and add it to the corresponding coefficients across all blocks. The perturbed image $\hat{\mathbf{x}}$ is obtained via Inverse BDCT (IBDCT) and projection:

$$\hat{\mathbf{x}} = \Pi_{\varepsilon,\mathbf{x}}\big(\text{IBDCT}(\mathbf{I}[c,:,i,j] + \hat{\mathbf{w}}_{c,i,j}, w)\big) \tag{10}$$

where $\Pi_{\varepsilon,\mathbf{x}}(\cdot)$ denotes the projection operator that maps the input onto a hypersphere centered at $\mathbf{x}$ with radius ε.

Observation: The Correlation with Clean Image Statistics. We applied BFS to 1,000 ImageNet validation images across diverse architectures. Results on ResNet-50, ViT-B-32, and adversarially trained WideResNet-50 (Fig. 3) reveal that model sensitivity (CE loss) exhibits an oscillatory decay from low to high frequencies. Crucially, we observe that this sensitivity pattern is closely related to the frequency-domain

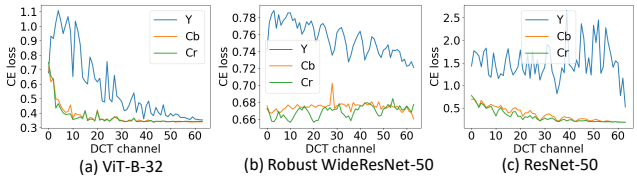


Figure 3: Classification sensitivity analysis of frequency bands via the proposed BFS. Higher CE loss indicates greater sensitivity of the corresponding BDCT frequency band.

variance of clean images. We compute the standard deviation of frequency coefficients across blocks for clean images as:

$$\sigma_{c,i,j} = \sqrt{\frac{1}{Z}\sum_{z=0}^{Z-1} |\mathbf{I}[c,z,i,j] - \mu_{c,i,j}|^2}. \tag{11}$$

As shown in Fig. 4, the distribution of $\sigma_{c,i,j}$ exhibits a similar decaying trend. To quantitatively assess this relationship, we compute the Pearson correlation between the channel-wise BFS sensitivity curves and the corresponding clean-image frequency variance profiles. As summarized in Table 1, we observe a strong and statistically significant positive correlation across standard architectures such as ViT-B-32 and ResNet-50. For adversarially trained models, this positive correlation remains evident in the luminance channel (Y) (e.g., $\bar{r} \approx 0.7$), while the chrominance channels exhibit more heterogeneous or mildly negative correlations. Importantly, this does not alter the overall trend, since the sensitivity of chrominance channels is consistently much lower than that of the luminance channel across all models, as reflected by their substantially smaller CE responses in Fig. 3.

While targeting model-specific sensitivity peaks in BFS curves (BFS-peak) may ideally seem optimal, identifying the exact band is almost infeasible in black-box settings where the architecture is unknown. Moreover, empirical evidence (details in the App. D.1 [41]) reveals that compared to clean-image statistics, BFS-peak yields only marginal gains on matched models while suffering severe degradation on mismatched ones. Therefore, we leverage the intrinsic frequency properties of clean images to establish a practical, model-agnostic prior. Importantly, BFS statistics are used only to motivate and validate this prior, not as a required attack-time component. For unknown victim models, the attacker can directly use the per-image clean variance prior; optional BFS curves can be estimated offline from public in-domain images with surrogate models, without accessing target models' private data.

4.3.2 Methodology: Constructing Initial Directions

Based on the observation above, we propose using the clean image frequency variance $\sigma_{c,i,j}$ directly as a prior to construct an initialization direction $\mathbf{d_n}$. For comparison and diversity,

Algorithm 1 DPAttack Process

Input: Target classifier $\mathcal{F}$, clean RGB image $\mathbf{x}$, GT label y, perturbation constraint ε, candidate block size set $\mathbb{W}$.
Output: Adversarial example $\mathbf{x}^*$.

1: Obtain optimal block size w and initialization direction $\mathbf{d_n}$ via Dynamic Blocksize Selection: $(w, \mathbf{d_n}, r) \leftarrow \mathrm{DBS}(\mathcal{F}, \mathbf{x}, y, \mathbb{W})$ {*refer to Alg. 2 in [41]*}
2: Initialize low-pass filtered direction $\phi(\hat{\mathbf{d}}_\mathbf{r})$ with $h \leftarrow w$.
3: Perform initial direction selection and boundary search:
4: $(\mathbf{d}_0, r_0) \leftarrow \mathrm{BiLiSearch}(\mathcal{F}, \mathbf{d_n}, \phi(\hat{\mathbf{d}}_\mathbf{r}), \mathbf{x}, y, r)$ {*refer to Alg. 3 in [41]*}
5: Set current best parameters: $\mathbf{d}_{\text{best}} \leftarrow \mathbf{d}_0$ and $r \leftarrow r_0$.
6: **while** Query Budget > 0 **and** $r > \varepsilon$ **do**
7: Update $\mathbf{d}_{\text{best}}$ and r via Pattern-Driven Optimization (Sec. 4.4).
8: **end while**
9: **return** $\mathbf{x}^* \leftarrow \mathrm{clip}(\mathbf{x} + r \cdot \mathbf{d}_{\text{best}}, [0,1])$

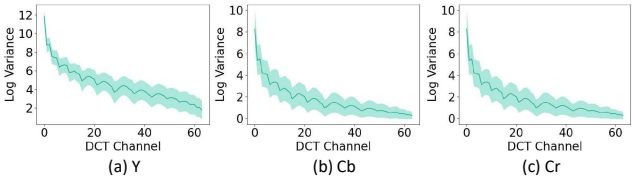


Figure 4: Clean image frequency statistics. The solid line and the shaded region represent the mean and standard deviation of the log-variance $\log(\sigma^2_{c,i,j} + 1)$ (defined in Eq. (11)).

Table 1: Pearson correlation analysis ($\bar{r}^{p-\text{value}}$) between the proposed BFS profile and the clean image frequency variance across different models and color channels.

Channel	ViT-B-32	WRS-50	ResNet-50
Y	$0.8074^{7.69E-16}$	$0.7024^{9.87E-11}$	$0.0327^{7.97E-01}$
Cb	$0.9274^{3.61E-28}$	$-0.1714^{1.76E-01}$	$0.9137^{6.18E-26}$
Cr	$0.8844^{3.44E-22}$	$-0.4087^{7.99E-04}$	$0.9618^{1.40E-36}$

[1] $\bar{r}$: the correlation coefficient; p-value: the statistical significance; WRS-50: Robust WideResNet-50.

we also design two alternative directions, $\mathbf{d_b}$ and $\mathbf{d_r}$, based on gradient local similarity [32] and color sensitivity [2].

1. Variance-Based Direction ($\mathbf{d_n}$). We construct the frequency perturbation matrix $\tilde{\Delta}$ by sampling noise coefficients $\mathbf{w}_{c,i,j} \sim \mathcal{N}(0, \sigma^2_{c,i,j})$ for each frequency band, where $\sigma^2_{c,i,j}$ is computed solely from the clean image $\mathbf{x}$ across its BDCT blocks at attack time. The spatial perturbation direction is then derived by:

$$\mathbf{x}'_\mathbf{n} = \mathrm{IBDCT}(\mathbf{I} + \tilde{\Delta}, w), \quad \mathbf{d_n} = \mathrm{sgn}(\mathbf{x}'_\mathbf{n} - \mathbf{x}), \quad \text{with } \mathrm{sgn}(0) := 1. \tag{12}$$

Zero elements in $\mathbf{d_n}$ are replaced with 1 to maintain a dense search space.

2. Gradient-Similarity-Based Direction ($\mathbf{d_b}$). Inspired by HRayS, we verify that gradient spatial correlations persist even after sign quantization (Fig. 5(a)). To leverage this sign-level coherence, we design $\mathbf{d_b}$ as a block-wise structure con-

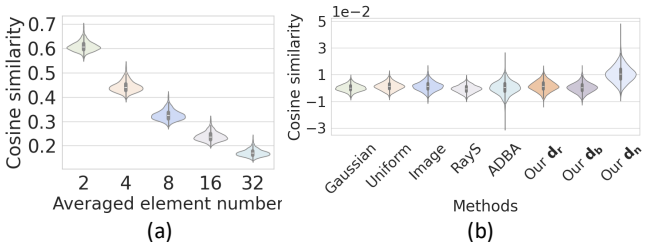


Figure 5: Cosine similarity between the true gradient sign and: (a) its block-wise averaged approximations; (b) various initialization directions.

sisting of $\hat{m}$ segments of length $\hat{n}$, where each segment is assigned a uniform sign:

$$\mathbf{d_b} = \left((-1)^{\lfloor (i-1)/\hat{n} \rfloor}\right)_{i=1}^{\hat{m}}. \tag{13}$$

3. Square-Color-Based Direction ($\mathbf{d_r}$). Leveraging the sensitivity of DNNs to square-shaped color perturbations [2], we generate $\mathbf{d_r}$ as the sign of the residual between the clean image $\mathbf{x}$ and a randomly square-perturbed image $\mathbf{x}'_r$, where the square has side length h. That is:

$$\mathbf{d_r} = \mathrm{sgn}(\mathbf{x}'_r - \mathbf{x}), \quad \text{with } \mathrm{sgn}(0) := 1. \tag{14}$$

Details for generating $\mathbf{x}'_r$ are presented in App. E of [41].

Low-Frequency Focusing Method (ϕ). To mitigate inter-band interference, we introduce a wrapper $\phi(\cdot)$ that confines perturbations to the low-frequency component $\mathbf{x}_{\mathrm{LL}}$ via Discrete Wavelet Transform (DWT) [23]. We prioritize DWT over BDCT to minimize hyperparameters, reserving BDCT for generating pixel-domain block-wise artifacts. We extract the low-pass component $\mathbf{x}_{\mathrm{LL}} \in \mathbb{R}^{C\times\hat{H}\times\hat{W}}$ via:

$$\mathbf{x}_{\mathrm{LL}}, \mathbf{x}_{\mathrm{others}} = \mathrm{DWT}(\mathbf{x}, dl), \tag{15}$$

where $\hat{H} = H/2^{dl}$ and $\hat{W} = W/2^{dl}$. The decomposition level dl is bounded by $\lfloor \log_2(\bar{w}) \rfloor$, where $\bar{w}$ is the pattern size of the underlying strategy (i.e., w for $\mathbf{d_n}$, $\hat{n}$ for $\mathbf{d_b}$, or h for $\mathbf{d_r}$). We isolate low frequencies by setting the residuals $\mathbf{x}_{\mathrm{others}} = \mathbf{0}$ and applying a perturbation $\hat{\mathbf{d}}_\mathbf{a} \in \{\hat{\mathbf{d}}_\mathbf{n}, \hat{\mathbf{d}}_\mathbf{b}, \hat{\mathbf{d}}_\mathbf{r}\}$ to $\mathbf{x}_{\mathrm{LL}}$. Note that $\hat{\mathbf{d}}_\mathbf{a}$ is generated with pattern size scaled by 2^{-dl} (e.g., $w/2^{dl}$ for $\hat{\mathbf{d}}_\mathbf{n}$). The perturbed low-frequency component $\hat{\mathbf{x}}_{\mathrm{LL}} = \mathbf{x}_{\mathrm{LL}} + \hat{\mathbf{d}}_\mathbf{a}$ is then projected back to the pixel space via Inverse DWT:

$$\mathbf{x}'_\mathbf{a} = \mathrm{IDWT}(\hat{\mathbf{x}}_{\mathrm{LL}}, \mathbf{0}, dl). \tag{16}$$

Encapsulating this process as $\phi(\hat{\mathbf{d}}_\mathbf{a})$, we derive the final direction $\mathbf{d_a}$ by taking the sign of the difference, ensuring non-zero entries:

$$\mathbf{d_a} = \mathrm{sgn}(\mathbf{x}'_\mathbf{a} - \mathbf{x}), \quad \text{with } \mathrm{sgn}(0) := 1. \tag{17}$$

Table 2: Comparison of curvature (κ) and perturbation distances (ℓ_∞, ℓ_2) across initialization methods. "Image" denotes a random sample from a different class.

Metrics	Curvature $\kappa\uparrow$		$\ell_\infty\downarrow$		$\ell_2\downarrow$	
Methods	Avg.	Med.	Avg.	Med.	Avg.	Med.
Gaussian	58,736	49,404	0.363	0.341	49.4	46.7
Uniform	58,273	48,932	0.315	0.318	49.7	48.6
Image	57,873	47,068	0.362	0.364	55.2	54.2
RayS	58,167	45,454	0.518	0.533	147.8	150.1
ADBA	39,408	33,592	0.354	0.339	114.9	121.2
Our $\mathbf{d_r}$	58,308	47,349	0.091	0.083	35.1	32.1
Our $\mathbf{d_b}$	60,934	50,192	0.108	0.072	39.3	27.2
Our $\mathbf{d_n}$	**68,502**	**54,743**	**0.078**	**0.069**	**28.9**	**26.1**

4.3.3 Validation and Dynamic Selection

We benchmark our initialization directions against standard baselines (Gaussian, Uniform, RayS, ADBA, and random images) using ResNet-50. The resulting starting points are evaluated quantitatively via three metrics: 1) Perturbation distance (ℓ_2/ℓ_∞); 2) Local curvature κ (indicating optimization speed);[3] and 3) Cosine similarity with the true gradient sign. As detailed in Table 2 and Fig. 5, $\mathbf{d_n}$ consistently demonstrates superior geometric properties: it achieves the highest gradient sign alignment (Fig. 5(b)) and significantly smaller boundary distances with higher local curvature compared to all baselines. These advantages are critical for optimization; indeed, our ablation study (Sec. 5.6) confirms that replacing $\mathbf{d_n}$ with other initializations severely degrades the attack success rate.

To theoretically justify this advantage, we model initialization as a sign-alignment problem. We show that, under a frequency-variance prior and an informative surrogate-gradient condition, sampling from a BDCT frequency distribution positively correlated with model sensitivity yields positive expected alignment with the true gradient sign up to a surrogate-transfer error $\varepsilon_{\mathrm{sur}}$ (as formalized in **Theorem 3**). This analysis is rooted in the Arcsine Law [52] (see proof in App. G of our arXiv version [41]), which connects positive value correlation to positive sign correlation for jointly Gaussian variables. In our setting, the frequency-variance prior biases $\mathbf{d_n}$ toward frequency bands that are, on average, more sensitive to the model; when this induced alignment dominates the transfer error ($\gamma > \varepsilon_{\mathrm{sur}}$), positive expected alignment follows. In contrast, isotropic random initialization distributes energy uniformly across frequencies and yields zero expected alignment in expectation. Under a local linear decision-boundary assumption, this positive alignment further implies a smaller expected boundary distance (as formalized in Corollary 1 of App. G.3 in [41]), explaining the empirical advantages of $\mathbf{d_n}$ in Table 2.

Dynamic Block Size Selection. While our theoretical anal-

[3] Optimization behaviors and κ estimation are detailed in Apps. H and I [41].

ysis supports the use of frequency-based initialization, empirical results (Sec. 5.6) indicate that the attack performance exhibits mild sensitivity to the choice of the BDCT block size for a given victim model. This sensitivity arises because the block size controls the granularity of the frequency decomposition, which in turn affects both the correlation between the initial direction and the true gradient sign, as well as the efficiency of the subsequent optimization process. To account for this variability without manual tuning, we introduce a Dynamic Block Size Selection (DBS) strategy that query-efficiently identifies a suitable block size w.

BiLiSearch Directions. Even with optimized block sizes, relying on a single initialization is not universally optimal against unknown defenses (e.g., adversarial training). While $\mathbf{d_n}$ is generally superior, we observe that the color-based alternative $\phi(\hat{\mathbf{d}}_\mathbf{r})$ yields marginal gains on adversarially trained models (see Table 9). To ensure robustness, we propose BiLiSearch, a hybrid strategy combining binary and line search to efficiently select the direction $\mathbf{d}_0$ with the minimal boundary distance r_0:

$$\mathbf{d}_0, r_0 = \text{BiLiSearch}(\mathbf{d_n}, \phi(\hat{\mathbf{d}}_\mathbf{r})), \quad \mathbf{x}_0^* = \text{clip}(\mathbf{x} + r_0\mathbf{d}_0, [0,1]). \tag{18}$$

Theorem 3 (Positive Expected Alignment under Frequency-Variance Prior)**.** *Let $\mathbf{g} = \nabla_\mathbf{x}\mathcal{L}(\mathbf{x}, y)$ and $\mathbf{u} = \text{sgn}(\mathbf{g}) \in \{-1,1\}^d$ denote the true gradient sign. Let the initialization be $\mathbf{x}_0 = U\mathbf{z}$, where $U \in \mathbb{R}^{d\times d}$ is the orthonormal BDCT basis and $\mathbf{z}$ is a zero-mean Gaussian vector with frequency-band variances $\{v_q\}_{q=1}^Q$. The initialization direction is $\mathbf{d_n} = \text{sgn}(\mathbf{x}_0)$. Let $\mathbf{h} = U\tilde{\mathbf{h}}$ be a Gaussian surrogate gradient signal with frequency-band variances $\{s_q\}_{q=1}^Q$. For $\rho_i = \text{Corr}((\mathbf{x}_0)_i, \mathbf{h}_i)$, let $\gamma := d^{-1}\sum_{i=1}^d \frac{2}{\pi}\arcsin(\rho_i)$. If $\gamma > 0$ and the surrogate-to-true-gradient transfer error is bounded by $\varepsilon_{\text{sur}} < \gamma$, then*

$$\frac{1}{d}\mathbb{E}[\langle \mathbf{d_n}, \mathbf{u}\rangle] \geq \gamma - \varepsilon_{\text{sur}} > 0. \tag{19}$$

In contrast, for isotropic random initialization $\mathbf{d}_{\text{rand}}$ that is coordinate-wise symmetric and independent of $\mathbf{u}$, we have $d^{-1}\mathbb{E}[\langle \mathbf{d}_{\text{rand}}, \mathbf{u}\rangle] = 0$.

4.4 Pattern-Driven Optimization (PDO)

Having established the initial direction $\mathbf{d}_0$, the next challenge is efficient direction refinement. Unlike random noise, $\mathbf{d}_0$-predominantly sourced from the BDCT-based $\mathbf{d_n}$-naturally exhibits strong spatial coherence in the pixel space, forming contiguous blocks of identical signs ("runs"), as illustrated in Figs. 2(a) and (c). However, SOTA hard-label attacks such as RayS and ADBA adopt rigid, pattern-agnostic dyadic bisection, which blindly partitions the pixel space and fragments these coherent structures. This mismatch disrupts the intrinsic geometry of $\mathbf{d}_0$ and leads to *gain cancellation* (formally defined in Lemma 5 of App. L [41]), where conflicting gradient signals within a partition offset each other, resulting in query inefficiency.

To address this issue, we propose the *Pattern-Driven Optimization (PDO)* module, which explicitly aligns the search process with the spatial patterns of $\mathbf{d}_0$. Instead of operating in the full pixel space of dimension d, PDO treats each sign-consistent run as an atomic search unit, effectively reducing the optimization problem to a lower-dimensional *pattern space* with dimensionality equal to the number of runs $M \ll d$. For example, on ImageNet ($d \approx 150$k), using block sizes $w \in \{4, 8, 16\}$ reduces M to approximately 20k, 11k, and 6k, respectively. By preserving spatial coherence and avoiding destructive partitioning, PDO enables more efficient gradient sign approximation that scales with the compressed pattern structure rather than the raw pixel dimensionality.

4.4.1 Algorithm Implementation

To operationalize this, PDO constructs a search tree T_{pat} aligned with the pattern of $\mathbf{d}_0$, via three phases:

Phase 1: Pattern-Driven Tree Construction.

- Sorting and indexing: Define a permutation Γ sorting $\mathbf{d}_0$ such that $\mathbf{d}_0[\Gamma(1)] \leq \cdots \leq \mathbf{d}_0[\Gamma(d)]$.
- Atomic partitioning: Partition Γ into atomic units $\{B_1, \ldots, B_M\}$ grouping contiguous indices, effectively capturing the "runs" in $\mathbf{d}_0$.
- Hierarchical grouping: Build a binary hierarchy where, at level $s \in [1, \lceil \log_2 M \rceil]$, atomic units are merged into 2^s groups $\{\mathcal{G}_j\}_{j=1}^{2^s}$ for coarse-to-fine search.

Phase 2: Sign Flipping and Direction Update. We refine $\mathbf{d}_{\text{best}}$ (initialized as $\mathbf{d}_0$) by iteratively flipping signs within adjacent group pairs $(\mathcal{G}_{2l-1}, \mathcal{G}_{2l})$ to generate candidates $\bar{\mathbf{d}}_1$ and $\bar{\mathbf{d}}_2$. To maximize efficiency, we employ a history-aware update rule:

$$\mathbf{d}_{\text{best}} = \begin{cases} \mathbf{d}_{\text{best}} & \text{if } \mathcal{F}(\mathbf{x}+r\bar{\mathbf{d}}_1) = y \;\&\; \mathcal{F}(\mathbf{x}+r\bar{\mathbf{d}}_2) = y \\ \Lambda(\bar{\mathbf{d}}_1, \bar{\mathbf{d}}_{\text{his}}) & \text{if } \mathcal{F}(\mathbf{x}+r\bar{\mathbf{d}}_1) \neq y \;\&\; \mathcal{F}(\mathbf{x}+r\bar{\mathbf{d}}_2) = y \\ \Lambda(\bar{\mathbf{d}}_2, \bar{\mathbf{d}}_{\text{his}}) & \text{if } \mathcal{F}(\mathbf{x}+r\bar{\mathbf{d}}_1) = y \;\&\; \mathcal{F}(\mathbf{x}+r\bar{\mathbf{d}}_2) \neq y \\ \Lambda(\bar{\mathbf{d}}_1, \bar{\mathbf{d}}_2) & \text{Otherwise.} \end{cases} \tag{20}$$

Here, Λ selects the direction with the smaller boundary distance. Notably, we use a buffer $\bar{\mathbf{d}}_{\text{his}}$ to cache solitary successful directions, preserving promising candidates that baseline methods (e.g., ADBA) would discard.

Phase 3: Refinement and Loop. We increment the granularity level s to progressively refine the search.[4] The complete workflow is visualized in Fig. 2(b).

[4] A fallback that splits atomic units or re-initializes $\mathbf{d}_0$ is only triggered in rare low-dimensional cases (e.g., CIFAR-10) and was not observed on higher-resolution datasets such as ImageNet.

4.4.2 Theoretical Guarantees

Having defined the algorithm, we now utilize a signal recovery framework to justify why aligning the search tree with initialization patterns (T_{pat}) outperforms standard dyadic trees (T_{dyad}) employed by HRayS and ADBA.

Theorem 4 (Per-query Sign Alignment Dominance)**.** *Let* $\mathbf{u} = \text{sgn}(\nabla_{\mathbf{x}}\mathcal{L})$ *be the true gradient sign, and let* $\mathbf{d}_t \in \{\pm 1\}^d$ *be the perturbation direction after* t *queries. We measure sign alignment using the agreement* $A(\mathbf{d}_t, \mathbf{u}) := \frac{1}{d}\sum_{i=1}^{d} \mathbb{I}\{\mathbf{d}_{t,i} = \mathbf{u}_i\}$*. Under the assumptions of local block coherence and a weakly positive initial correlation, the expected agreement* A_t *after* t *queries satisfies:*

$$\mathbb{E}\left[A_t^{\text{pat}}\right] \geq \mathbb{E}\left[A_t^{\text{dyad}}\right], \quad \forall t \in \mathbb{N}, \tag{21}$$

where pat *and* dyad *denote the pattern-driven strategy and the blind dyadic baseline, respectively.*

Proof in App. L of [41]. The inequality degenerates to an equality only if $\mathbf{d}_0$ lacks spatial variation (e.g., $\mathbf{d}_0 = \mathbf{1}$) or aligns perfectly with the dyadic grid. We provide empirical verification of this dominance in Sec. 4.4.3. This per-step advantage scales into a significant improvement in asymptotic query complexity, as formalized in **Theorem 5** (see App. M in [41] for full proof). Throughout Theorem 5, we use the number of tree expansions as a proxy for query complexity. This is justified because, in RayS-style hard-label attacks (including our PDO), each node expansion incurs at most a constant number of model queries. Therefore, the asymptotic bounds in Theorem 5 translate directly to query complexity up to constant factors. Also, the complexity T quantifies the cost for *full* sign recovery. In practice, attacks succeed upon achieving *sufficient alignment* (partial recovery); thus, actual consumption is usually lower than T. We attribute this complexity gap to a structural phenomenon we term *Intrinsic Gradient Shattering*, a detailed empirical analysis of which is provided in the following validation.

Theorem 5 (Query Complexity under Block Sign-Coherence)**.** *Let the true gradient sign* $\mathbf{u}$ *consist of* K *spatially coherent blocks* $\{B_k\}_{k=1}^{K}$*. To identify descent directions aligned with* $\mathbf{u}$*, the expected query complexities for* dyadic *search* (T_{dyad}) *and* pattern-driven *search* (T_{pat}) *satisfy:*

$$T_{\text{dyad}} = \Omega\left(\sum_{k=1}^{K} \log_2 \frac{d}{|B_k|}\right), \quad T_{\text{pat}} = O\left(\sum_{k=1}^{K} \gamma_k\right), \tag{22}$$

where $\Omega(\cdot)$ *and* $O(\cdot)$ *denote asymptotic lower and upper bounds.* γ_k *is the number of* $\mathbf{d}_0$*-runs intersecting block* B_k*. In particular, the pattern-driven strategy is asymptotically more efficient in expectation when* $\gamma_k < \log_2(d/|B_k|)$ *on average.*

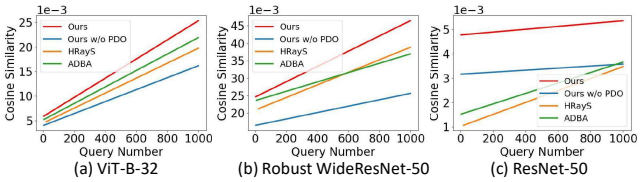


Figure 6: Evolution of cosine similarity between the true gradient sign and the perturbation direction.

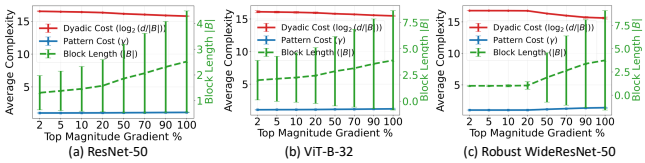


Figure 7: Empirical validation of Theorem 5. Results for additional models are provided in Fig. 15 in [41].

4.4.3 Empirical Validation

To empirically verify Theorems 4 and 5, we track the cosine similarity between the search direction $\mathbf{d}_{\text{best}}$ and the true gradient sign $\text{sgn}(\nabla\mathcal{L}(\mathbf{x}, y))$ (averaged over 1,000 ImageNet images). The results are visualized in Fig. 6. Note that to improve clarity, curves undergo a linearly fitted process, which causes a minor visual divergence at $t = 0$ between "Ours" and "Ours w/o PDO" (using standard dyadic split), even though they originate from the exact same initial direction. The observed trends strongly confirm our theoretical claims:

1) The "Ours" curve consistently upper-bounds the baseline "Ours w/o PDO" across architectures. This empirical gap confirms Theorem 4, demonstrating that incorporating initial pattern priors generally yields a better descent direction for any given query budget t.

2) The substantial gap between the slow ascent of "Ours w/o PDO" and the steep rise of "Ours" also confirms the complexity divergence predicted by Theorem 5. We attribute this to *Intrinsic Gradient Shattering*—a phenomenon where top-magnitude gradient components, critical for optimization, exhibit rapid spatial oscillation. As visualized by the green lines in Fig. 7, the average sign-consistent block size $|B|$ for these top-2% magnitude gradients collapses to ≈ 1 in CNNs and ≈ 2 in ViTs. This inherent fragmentation triggers two distinct regimes: (i) **Dyadic Collapse:** The rigid partitioning of dyadic strategies is forced into a worst-case scenario by these pixel-level oscillations. This drives the search depth toward the theoretical limit ($\approx \log_2 d$), incurring a query complexity of $\approx 16 \sim 17$ per block—matching the value of dyadic costs in Fig. 7. (ii) **Pattern Resilience:** In contrast, PDO sidesteps this shattering by adaptively aligning with the run-length boundaries of $\mathbf{d}_0$. By treating these oscillations as coherent atomic units rather than disjoint pixels, PDO maintains a near-optimal fragmentation factor $\gamma \approx 1$ (represented by the blue

lines in Fig. 7), significantly accelerating convergence.
3) In Fig. 6, our method consistently outperforms HRayS and ADBA. Additionally, the increasing HRayS trend validates Theorem 2, confirming that while less efficient than our method, the standard dyadic hierarchy still guarantees a progressive reduction in approximation error.

5 Evaluation

5.1 Experimental Setup

Datasets and Victim Models. We evaluate our method across a diverse range of tasks. For standard classification on CIFAR-10 [36] and ImageNet-1K [17], we target representative architectures (e.g., ResNet-18/50, ViT-B-32, DenseNet-121) alongside adversarially trained WideResNet variants [14, 20]. We also assess the ViT-H-14 CLIP [31] foundation model and test robustness against context shifts on ObjectNet [6]. For dense prediction, we employ MS-COCO [40] (attacking Faster-RCNN [45]) and SA-1B [35] (attacking SAM [35]). Furthermore, we include ImageNet-C [28] and PathMNIST [60] to evaluate performance on corrupted and non-natural (medical) domains using their respective top-performing models. See App. A for the full list of over 15 evaluated architectures and image sizes for each dataset.
Metrics. We evaluate attack effectiveness using Attack Success Rate (ASR) within a query budget (Max.Q). We compute ASR$= \frac{1}{N}\sum_{i=0}^{N-1} \mathbb{I}(\mathcal{F}(\mathbf{x}_i^*) \neq y_i)$ and set $N = 200$ following [55]. To assess efficiency, we further report the average (Avg.Q) and median (Med.Q) queries among successful attacks.
Comparative Methods. We evaluate two variants of our method: (1) Ours$_{\text{opt}}$, which uses the fixed block size that achieves the highest average ASR across query budgets, and (2) Ours$_{\text{dyn}}$, which employs our dynamic block size selection strategy. We compare them against SOTA query-based hard-label attacks: TtBA [57], HRayS [9], ADBA [56], HSJA [8], and Bounce [55] with both random (BouR) and targeted (BouT) initializations. For ℓ_2 comparisons, we include SurFree [43], Tangent [42], and CGBA(-H) [48]. Implementation details, including the setting of decomposition level dl and hyperparameter configurations, are provided in App. A.

5.2 Performance on Standard Classifiers

We evaluate the efficacy of DPAttack against standard classifiers trained on CIFAR-10, ImageNet-1K, ImageNet-C and PathMNIST. Primary results are summarized in Tables 3, 10-17, where extensive evaluations across diverse architectures and domains are detailed. **ASR.** Across nearly all scenarios, our method consistently outperforms SOTA baselines. The gap is particularly pronounced under tight query budgets. For instance, on ResNet-50 with Max.Q$=$ 50, our$_{\text{dyn}}$ surpasses the SOTA ADBA by 32.5% in ASR while consuming only approximately two-thirds of its Avg.Q and Med.Q. Note that for extremely low budgets (Max.Q$<$ 20), the ASR of Ours$_{\text{dyn}}$ drops due to the DBS warm-up overhead, which costs about 15/25 queries per image on standard/robust models; this can be mitigated by tuning the block size on a few images first and then using the most frequently selected block size for subsequent attacks, although per-image DBS remains preferable under larger budgets. **Query Efficiency.** DPAttack achieves competitive or lower query consumption compared to prior works at most cases—significantly outperforming the runner-up ADBA—while maintaining substantially higher ASR. Although our Avg.Q on ImageNet occasionally exceeds that of BouT, this stems from our significantly higher ASR: we successfully crack "hard" samples that require more queries, whereas baseline methods simply fail on them (yielding no query count contribution). **Additional Results.** Comparisons under varying perturbation magnitudes and ℓ_2 constraints are provided in Appendix C.1 and Appendix D, respectively. These results consistently demonstrate our method's superior ASR and query efficiency over SOTA baselines; notably, Ours$_{\text{dyn}}$ excels in ℓ_2 scenarios, often significantly surpassing the Ours$_{\text{opt}}$ version.

5.3 Performance on CLIP and APIs

CLIP. CLIP model is a versatile and prevailing foundation vision-language model for many downstream tasks and is suitable for zero-shot image classification, which achieves 77.4% Top-1 classification accuracy on ObjectNet. We evaluate our method against the CLIP model on ObjectNet to test its zero-shot robustness. To strictly simulate a hard-label black-box setting, we assume the adversary can only access the index of the predicted category (the class with the highest cosine similarity), without access to the specific similarity values or gradients. As shown in Table 4, our method demonstrates superior ASR and query efficiency compared to SOTA baselines. Notably, under a tight query budget (Max.Q$=$ 50), Ours$_{\text{dyn}}$ outperforms ADBA by 32% in ASR while halving the queries. Even with a generous budget ($> 3,000$), while the ASR gap narrows ($< 10\%$), our method remains significantly more efficient, reducing Med.Q by approximately 80% and Avg.Q by 40% compared to ADBA.
APIs. We further validate our approach on four popular commercial image labeling services: Imagga [33], Google Cloud Vision [24], Tencent DetectLabel Pro [50], and Baidu General Image Recognition [4]. The test images are randomly sampled from the ImageNet validation set. Due to the high cost of commercial API calls, we compare exclusively against ADBA, the strongest baseline identified in previous experiments. The results, summarized in Table 5, indicate that our algorithm consistently outperforms ADBA across all APIs and perturbation limits. For instance, on the Tencent API with $\ell = 0.1$, Ours$_{\text{opt}}$ achieves an ASR improvement of over 44% while consuming significantly fewer queries. Our method drastically reduces the query complexity from the tens or hun-

Table 3: The untargeted attack performance on ImageNet (ResNet-50) under $\ell_\infty = 0.05$ constraint.

Metrics	ASR								Avg.Q								Med.Q							
Max$Q$	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**
5	0.0	0.0	3.5	0.0	2.0	0.0	**20.5**	0.0	0.0	-	5	-	5	-	5	-	-	-	5	-	5	-	5	-
10	0.0	0.0	9.5	0.0	2.0	0.0	**37.0**	-	-	-	6		10	-	6	-	-	-	6	-	10	-	5	-
20	6.5	6.5	9.5	0.0	2.0	3.5	37.0	**39.5**	14	14	6	-	20	17	6	19	14	14	6	-	20	17	5	19
50	6.0	6.0	17.0	0.0	16.5	22.5	51.5	**55.0**	13	13	20	-	42	32	19	23	13	13	6	-	40	31	13	20
80	14.0	14.0	17.5	6.5	21.0	38.0	57.0	**64.0**	40	40	21	75	52	46	24	29	59	59	6	75	56	42	15	20
100	15.0	16.0	16.0	7.5	25.0	49.5	62.0	**66.0**	40	37	23	77	60	56	29	31	59	38	6	75	57	59	15	20
200	17.5	17.5	22.0	10.5	51.0	71.0	**75.5**	75.0	75	64	72	87	109	87	49	46	99	99	48	75	104	80	26	22
300	20.0	18.5	21.0	14.0	54.0	79.5	**83.5**	**83.5**	77	71	54	126	136	104	67	66	99	99	6	82	121	84	34	22
400	16.5	16.0	23.5	16.0	69.0	85.0	88.5	**89.0**	61	60	72	158	160	117	83	85	99	99	7	144	143	85	38	24
500	20.5	18.0	25.5	19.0	70.5	87.5	90.0	**92.0**	92	81	85	184	198	128	88	97	99	99	88	146	174	88	38	29
1k	20.5	18.5	22.0	32.5	89.5	93.5	95.5	**97.5**	164	110	184	414	312	163	127	131	150	149	8	404	226	97	44	36
3k	23.0	19.5	23.0	59.5	97.5	99.0	99.5	**100.0**	352	209	196	1,117	431	236	199	180	150	149	139	965	252	100	48	39
5k	21.0	19.0	22.5	67.5	**100.0**	**100.0**	**100.0**	**100.0**	210	140	319	1,441	522	249	222	180	149	149	138	978	263	102	49	39

Table 4: The untargeted attack performance on ObjectNet (ViT-H-14-CLIPA-336) under $\ell_\infty = 0.05$ constraint.

Metrics	ASR								Avg.Q								Med.Q							
Max$Q$	HSJA	BouR	BouT	TtBA	HRayS	ADBA	Ours$_{opt}$	Ours$_{dyn}$	HSJA	BouR	BouT	TtBA	HRayS	ADBA	Ours$_{opt}$	Ours$_{dyn}$	HSJA	BouR	BouT	TtBA	HRayS	ADBA	Ours$_{opt}$	Ours$_{dyn}$
10	0.0	0.0	10.0	0.0	0.0	0.0	**21.5**	0.0	-	-	6	-	-	-	5		-	-	5	-	-	-	5	
50	5.5	5.5	15.0	0.0	5.0	7.5	37.0	**39.5**	13	13	20	-	41	41	16	22	12	12	6	-	39	42	10	20
100	20.5	18.0	22.5	9.5	13.0	20.5	44.5	**50.0**	40	42	35	77	72	61	25	33	61	58	6	76	81	58	15	20
500	23.5	23.5	23.0	22.0	38.0	49.5	65.5	**67.5**	95	91	72	207	178	156	98	92	102	98	88	190	157	118	34	28
1k	25.0	23.5	25.5	32.5	51.0	61.0	77.0	**78.5**	136	129	137	363	326	257	194	186	151	149	6	313	220	162	55	49
3k	26.5	22.5	26.5	58.5	73.0	82.5	**88.5**	**88.5**	242	141	202	1,105	770	636	371	364	152	149	137	850	430	290	86	65
6k	27.0	25.0	29.5	73.0	82.0	88.0	**92.5**	**92.5**	637	130	197	1,692	1,171	872	536	523	152	149	8	1,356	565	313	98	82
15k	33.0	24.5	27.5	83.0	91.5	95.5	**98.5**	98.0	1,123	412	137	2,437	2,135	1,527	1,039	1,022	154	148	7	1,650	853	362	124	92
30k	37.5	22.5	28.5	89.5	96.5	99.0	99.0	**99.5**	3,880	107	203	3,498	3,024	2,244	1,366	1,399	155	148	138	1,952	940	509	171	99
40k	40.0	24.5	30.0	89.5	95.5	99.5	99.5	**100.0**	6,552	114	225	4,460	1,211	2,426	1,367	1,555	153	149	8	2,125	651	518	141	103

dreds required by ADBA to merely $10-20$. This significant reduction in attack cost exposes the severe vulnerability of online APIs to low-budget adversaries, highlighting the urgent need for more robust defenses.

5.4 Performance on Defensive Methods

We evaluate our attack against both static (adversarial training) and dynamic stateful defense mechanisms (Blacklight).

Adversarial Training. We test attack effectiveness on robust WideResNet models (CIFAR-10 and ImageNet). As summarized in Table 6, our method achieves the highest ASR with superior query efficiency. Notably, we outperform HRayS by over 6% ASR on ImageNet. In contrast, ADBA shows limited improvement (0.5% on ImageNet) and even degradation (-0.5% on CIFAR-10) versus HRayS. Additional results under varying constraints are detailed in App. C.3.

Stateful Detection. We extend our evaluation to Blacklight [39], a SOTA defense that detects attacks by identifying highly correlated queries via hashing-based fingerprinting. Since Blacklight flags attacks when hash collisions within a sliding window exceed a threshold, we propose a randomized variant of DPAttack to break this correlation. Specifically, we inject Gaussian noise ξ into the update direction: $\mathbf{x}+r(\text{clip}(\bar{\mathbf{d}}_1+\xi,-\mathbf{1},\mathbf{1}))$. This stochasticity sufficiently diversifies the query patterns, preventing the generation of identical fingerprints and thereby evading the collision-based detection mechanism. Adhering to the official settings for ImageNet, we use a detection threshold of 25 and a window size of 50. Any attack triggering this threshold is marked as a failure,

Table 5: The untargeted attack performance on real-world APIs under different perturbation constraints. Abbreviations: ASR: A., Med.Q: M.Q, Avg.Q: A.Q.

ℓ_∞	APIs	Imagga			Google			Tencent			Baidu		
	Methods	A.	M.Q	A.Q	A.	M.Q	A.Q	A.	M.Q	A.Q	A.	M.Q	A.Q
0.05	ADBA	83	61	105	53	26	28	33	25	26	65	25	26
	Ours$_{opt}$	90	**6**	38	91	22	23	67	**10**	**15**	**88**	**9**	**13**
	Ours$_{dyn}$	**100**	23	**20**	**100**	**21**	**19**	**75**	23	20	**90**	21	20
0.1	ADBA	99	41	70	68	24	26	56	33	35	80	24	27
	Ours$_{opt}$	**100**	**4**	**8**	94	7	14	**100**	**14**	**14**	98	**7**	**14**
	Ours$_{dyn}$	**100**	16	16	**100**	**15**	**16**	97	17	16	**99**	16	16

Table 6: The untargeted attack performance on adversarially trained WideResNet. Max.Q: 1,000.

Dataset	Metrics	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**
CIFAR-10	ASR	4.5	3.5	3.5	1.5	20.5	20.0	**22.5**	21.0
	Avg.Q	140	141	58	361	290	211	234	205
	Med.Q	140	140	8	398	222	101	143	83
ImageNet	ASR	10.5	10.5	11.0	3.5	44.5	45.0	**50.5**	48.5
	Avg.Q	112	111	84	302	189	143	143	191
	Med.Q	149	146	6	223	91	58	73	90

and the query count before detection is recorded as 1$^{\text{st}}$DQ. As shown in Table 7, this strategy is remarkably effective: our method yields a 0% Detection Rate (Det$R$). In sharp contrast, competing SOTA methods are rendered ineffective, triggering the defense with a 100% Det$R$ (except BouT at 95%). We further compare with the randomized certifiable attack CA [29] (the surrogate-free binary-search variant) in App. E; under the same settings, DPAttack reaches the same ASR/Det$R$ while requiring smaller perturbations and far fewer queries.

Table 7: The untargeted attack performance on Blacklight against ViT-B-32. Max.Q: 100. Outside/inside parentheses: DPAttack with/without Gaussian-noise injection.

Metrics	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours**$_{\text{opt}}$	**Ours**$_{\text{dyn}}$
ASR↑	0	0	5	0	0	0	41 (22)	**43** (24.5)
Avg.Q↓	-	-	5	-	-	-	**39** (11)	40 (17)
Med.Q↓	-	-	5	-	-	-	31 (10)	**26** (17)
Det$R$↓	100	100	95	100	100	100	**0** (78)	**0** (76.5)
1$^{\text{st}}$DQ↑	9	9	8	11	5	2	- (15.7)	- (20.3)

Table 8: The untargeted attack performance on dense prediction tasks. Max.Q: 50.

Methods	Object Detection			Segmentation		
	ASR	Avg.Q	Med.Q	ASR	Avg.Q	Med.Q
ADBA	17.0	30	29	4.5	36	35
Ours $\mathbf{d_n}$	**43.5**	**18**	**6**	**36.0**	**14**	**6**
Ours$_{\text{opt}}$	35.5	16	10	35.5	17	10
Ours$_{\text{dyn}}$	39.5	22	20	33.5	20	18

Table 9: The ablation study on the initialization methods and PDO module. Abbreviation: WideResNet-50 (WRS-50); all one vector (**1**); dyadic search used in ADBA (ADBAS), see Step 2 of ADBA in Sec. 2.3. Max.Q: 500.

Model	Search	Gauss.	Unif.	Img.	**1**	$\mathbf{d_b}$	$\mathbf{d_r}$	$\mathbf{d_n}$	DDM
ViT	ADBAS	52.5	62.5	76.5	75.5	63.5	75.5	75.5	77.0
	PDO	70.5	72.0	76.0	75.5	82.5	81.5	**85.5**	**85.5**
WRS-50	ADBAS	16.5	18.0	36.5	45.0	27.5	35.0	28.5	37.0
	PDO	42.0	43.5	43.0	45.0	44.0	48.0	47.0	**50.5**

5.5 Generalization to Dense Prediction Tasks

To evaluate the generalization of DPAttack across diverse vision tasks, we evaluate its performance on object detection and segmentation against the runner-up, ADBA (detailed attack objectives in App. A). As summarized in Table 8, our method demonstrates superior efficacy and efficiency in both tasks. Notably, for segmentation, Ours $\mathbf{d_n}$ outperforms ADBA by over 31.5% in ASR. This significant margin validates that our frequency-based priors are highly effective even for complex dense prediction models. App. Table 15 shows that our $\mathbf{d_n}$ still sustains an ASR improvement of over 31% (OD) and 24.5% (Segmentation) with a larger 1,000-query budget.

5.6 Ablation Study

DDM. To evaluate our DDM initialization, we compare it against standard baselines (Gaussian, Uniform, Random Images, and All-ones vector **1**) while keeping the subsequent PDO search fixed. As shown in Table 9, DDM significantly outperforms traditional priors. For instance, on ViT, DDM achieves an ASR of 85.5%, outstripping Gaussian and Uniform initializations by 15.0% and 13.5%, respectively. On the robust WideResNet-50, DDM still provides a substantial gain (+8.5% over Gaussian). Notably, the advantage of DDM over individual components ($\mathbf{d_b}, \mathbf{d_r}, \mathbf{d_n}$) highlights the necessity of our integrated initialization pipeline. These results confirm that DDM provides a *structurally informative prior* that aligns with the target model's spectral sensitivity, thereby boosting search efficiency. Complementary studies on the individual contributions of $\mathbf{d_n}/\phi(\hat{\mathbf{d}}_\mathbf{r})$ in DDM are detailed in App. B. Results show that using our variance-based $\mathbf{d_n}$ alone can even yield higher ASR on standard models compared to DDM. Specifically, $\mathbf{d_n}$ is critical for enhancing overall ASR, while $\phi(\hat{\mathbf{d}}_\mathbf{r})$ balances performance across diverse models.

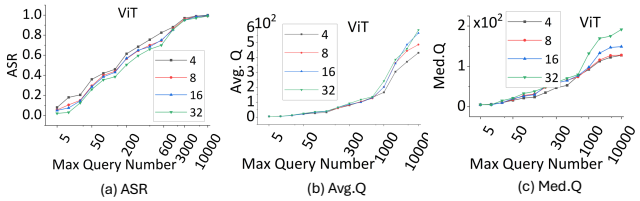


Figure 8: The influence of block size w used in BDCT.

PDO. The PDO module is designed to refine the initial direction. To assess its impact, we replace PDO with the basic structured dyadic search (denoted as ADBAS, the search strategy used in ADBA). Focusing on the last column in Table 9, we observe that under DDM initialization, PDO outperforms ADBAS by 8.5% on ViT and 13.5% on WideResNet-50. This empirically validates Theorems 4 and 5, proving that PDO's sign-consistent run partitioning is superior to fixed dyadic splitting. Moreover, PDO consistently enhances ASR across virtually all initialization methods.

Block Size w**.** We evaluate the sensitivity of the block size $w \in \{4, 8, 16, 32\}$ on ImageNet. As illustrated in Fig. 8, while the overall ASR remains relatively stable, the choice of w leads to fluctuations in query efficiency. Furthermore, the optimal w varies across architectures and settings. For instance, on ImageNet, WideResNet-50 and ViT-B-32 peak at $w = 16$ and $w = 4$, respectively, while ResNet-18 favors $w = 2$ on CIFAR-10. This "one-size-does-not-fit-all" observation underscores the necessity of our dynamic block size selection algorithm, which adaptively identifies a suitable w for a given target model to achieve better ASR and query efficiency.

6 Discussion

Targeted attacks are inherently more challenging than untargeted attacks due to their narrower adversarial regions and higher-curvature decision boundaries [48]. Although our BDCT initialization effectively captures frequency-band sensitivity and block-wise structural priors, it does not explicitly preserve Fourier phase information, which may limit its ability to guide perturbations toward a specific target class. A promising future direction is to combine the BFS amplitude prior with target-image Fast Fourier Transform phase information.

Regarding perceptibility, our method achieves higher ASR

while maintaining comparable Peak Signal-to-Noise Ratio, Structural Similarity Index Measure, and Learned Perceptual Image Patch Similarity values to baseline methods under the ℓ_2 constraint. Under the ℓ_∞ setting, the visual quality of our perturbations is similar to other sign-flip-based methods such as RayS and ADBA despite substantially improved ASR. Compared with other baselines, our perturbations introduce only slight additional visual degradation while providing significantly stronger attack effectiveness.

7 Conclusion

In this work, we propose DPAttack, a highly efficient hard-label black-box attack framework that bridges spectral sensitivity with spatial structural search. By leveraging a systematic analysis of adversarial sensitivity across BDCT frequency coefficients, we develop a customized initialization strategy that provides a structurally informative starting point. To exploit this prior, we introduce a pattern-driven direction optimization module, which significantly accelerates gradient sign estimation by searching along sign-consistent runs. Furthermore, we establish a novel theoretical framework for hard-label sign-flipping attacks, providing proofs for the superiority of our initialization and search strategies. Extensive experiments across various architectures, datasets, tasks and real-world APIs demonstrate that DPAttack consistently surpasses SOTA methods in both attack success rate and query efficiency.

Acknowledgments

We are grateful to the anonymous reviewers for their valuable guidance and insightful comments. J. Zhou's, F. Li's, and part of J. Liu's work was supported in part by Macau Science and Technology Development Fund under 001/2024/SKL, 0119/2024/RIB2 and 0110/2025/R1B2; in part by Research Committee at University of Macau under MYRG-CRG2025-00031-FST and MYRG-GRG2025-00086-FST; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515012536. I. Echizen's and part of J. Liu's work was supported by JSPS KAKENHI Grant JP24H00732; by JST CREST Grants JPMJCR20D3 and JPMJCR2562 including AIP challenge program; and by JST K Program Grant JPMJKP24C2 Japan.

Ethical Considerations

We structure our ethical analysis based on the stakeholder framework and principles outlined in the Menlo Report [5] and the USENIX Security Ethics Guidelines. Our research involves the development of a query-efficient hard-label black-box attack. We have identified the following key stakeholders: (1) Commercial Machine Learning as a Service (MLaaS) providers (Google, Baidu, Tencent, Imagga); (2) Authors of academic defense models (e.g., Blacklight); (3) Human subjects in the datasets; and (4) The broader research community and society.

Interaction with Commercial APIs (MLaaS Providers). To evaluate the practical efficiency of our attack and the robustness of real-world APIs, we tested our method against live APIs provided by Google, Baidu, Tencent, and Imagga.

- **Minimizing harm (beneficence):** We strictly limited the query volume to the minimum necessary to validate the attack's capability. Unlike Denial-of-Service (DoS) attacks or resource exhaustion exploits, our method focuses on finding adversarial examples with a low query budget. We maintained a conservative request rate to minimize the risk of degrading service quality for other users or incurring high computational costs.
- **Terms of Service (ToS) and legal compliance:** We acknowledge that probing APIs for vulnerabilities or adversarial robustness may technically violate certain standard ToS regarding reverse engineering or automated testing. However, consistent with the standard practice in security research, we deemed this limited violation necessary to understand the vulnerability of real-world systems to hard-label attacks. We utilized legitimately registered accounts and, where applicable, paid the required fees for API usage, ensuring no financial loss to the providers. We did not attempt to exfiltrate model weights, proprietary data, or compromise the underlying infrastructure.

Responsible Disclosure. Regarding the academic defense model *Blacklight*, we plan to share our detailed analysis with the authors upon publication. For the commercial APIs, since our attack demonstrates a general susceptibility of neural networks to adversarial examples rather than a specific implementation bug or infrastructure vulnerability (e.g., memory corruption), a standard Coordinated Vulnerability Disclosure process is less applicable. However, we are in the process of notifying the security teams of the respective vendors about the general robustness implications found in our study upon the acceptance of this work.

Dataset Usage and Privacy (Respect for Persons). Our experiments utilize several public datasets.

- **Biomedical data (PathMNIST):** We treated the PathMNIST dataset with high caution given its biomedical nature. We utilized only the publicly available of the dataset released by the original authors. We did not attempt to deanonymize any data subjects or use the data for any purpose other than evaluating algorithmic effectiveness.
- **Large-scale data:** For datasets containing human figures (e.g., SA-1B, MS-COCO, ImageNet), we adhered to their respective licenses. Our research focuses on pixel-level perturbations and does not generate offensive content or violate the dignity of the individuals depicted.

Risk-Benefit Analysis and Mitigation.

- **Potential misuse:** We recognize that publishing efficient hard-label attack code could potentially be used by malicious actors to evade content filters or integrity checks.
- **Justification for publication:** However, hard-label attacks represent a realistic threat model where attackers have limited access. By demonstrating that SOTA defenses, commercial APIs and foundation models are still vulnerable to query-efficient attacks, our work highlights a critical security gap. The benefits of motivating stronger defenses and exposing the limitations of current "security through obscurity" (reliance on label-only access) outweigh the risks.
- **Mitigation:** To mitigate potential harm, we will release our code with a responsible usage license.

Team Wellbeing. Our research team was not exposed to psychologically disturbing content during the handling of standard vision datasets. Furthermore, all API testing was performed using dedicated research infrastructure to ensure operational security and isolate experimental traffic.

Open Science

In alignment with the USENIX Security Open Science Policy and our commitment to reproducibility, we make all research artifacts associated with this paper publicly available. These artifacts, including data and code, can be accessed via the following repository: https://doi.org/10.5281/zenodo.20322560. The repository contains the datasets used in our evaluation, the source code for the attack algorithms and experimental analysis, as well as the corresponding model weights necessary to reproduce our results.

References

[1] Abdullah Al-Dujaili and Una-May O'Reilly. Sign bits are all you need for black-box attacks. In *ICLR*, 2020.

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *ECCV*, 2020.

[3] Apple. Apple core ml. https://developer.apple.com/cn/machine-learning/core-ml/, 2023.

[4] Baidu. Image recognition. https://cloud.baidu.com/doc/IMAGERECOGNITION/index.html, 2024.

[5] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. The menlo report. *IEEE S&P*, 10(2):71–75, 2012.

[6] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *NeurIPS*, volume 32, 2019.

[7] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *ICLR*, 2018.

[8] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *IEEE S&P*, 2020.

[9] Jinghui Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack. In *KDD*, 2020.

[10] Yiting Chen, Qibing Ren, and Junchi Yan. Rethinking and improving robustness of convolutional neural networks: a shapley value-based approach in frequency domain. *NeurIPS*, 2022.

[11] Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, JinFeng Yi, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In *ICLR*, 2019.

[12] Minhao Cheng, Simranjit Singh, Patrick Chen, Pin Yu Chen, Sijia Liu, and Cho Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. In *ICLR*, 2020.

[13] Minhao Cheng, Huan Zhang, Cho-Jui Hsieh, Thong Le, Pin-Yu Chen, and Jinfeng Yi. Query-efficient hard-label black-box attack: An optimization-based approach. In *ICLR*, 2019.

[14] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *NeurIPS*, 2021.

[15] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *ICML*, 2020.

[16] Media Cybernetics. Mediacy. https://mediacy.com/.

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

[18] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019.

[19] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018.

[20] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness. https://github.com/MadryLab/robustness, 2019.

[21] N Benjamin Erichson, Soon Hoe Lim, Francisco Utrera, Winnie Xu, Ziang Cao, and Michael W Mahoney. Noisymix: Boosting robustness by combining data augmentations, stability training, and noise injections. In *AISTATS*, 2024.

[22] Zheng Fang, Tao Wang, Lingchen Zhao, Shenyi Zhang, Bowen Li, Yunjie Ge, Qi Li, Chao Shen, and Qian Wang. Zero-query adversarial attack on black-box automatic speech recognition systems. In *ACM CCS*, pages 630–644, 2024.

[23] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Pearson, 4th edition, 2018.

[24] Google. Google cloud vision api. https://cloud.google.com/vision?hl=en.

[25] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *ICLR*, 2019.

[26] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *ICLR*, 2018.

[27] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021.

[28] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

[29] Hanbin Hong, Xinyu Zhang, Binghui Wang, Zhongjie Ba, and Yuan Hong. Certifiable black-box attacks with randomized adversarial examples: breaking defenses with provable confidence. In *ACM CCS*, pages 600–614, 2024.

[30] Christian Szegedy Ian J. Goodfellow, Jonathon Shlens. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[31] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. doi:10.5281/zenodo.5143773.

[32] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *ICLR*, 2019.

[33] Imagga. Ai-powered image tagging api. https://imagga.com/solutions/auto-tagging.

[34] Shuaifan Jin, He Wang, Zhibo Wang, Feng Xiao, Jiahui Hu, Yuan He, Wenwen Zhang, Zhongjie Ba, Weijie Fang, Shuhong Yuan, et al. Defending deep learning-based privacy attacks with gradient descent-resistant features in face recognition. In *USENIX*, 2024.

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023.

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

[37] Fengpeng Li, Kemou Li, Haiwei Wu, Jinyu Tian, and Jiantao Zhou. DAT: Improving adversarial robustness via generative amplitude mix-up in frequency domain. In *NeurIPS*, pages 127099–127128, 2024.

[38] Fengpeng Li, Kemou Li, Haiwei Wu, Jinyu Tian, and Jiantao Zhou. Toward robust learning via core feature-aware adversarial training. *IEEE TIFS*, 20:6236–6251, 2025.

[39] Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y Zhao. Blacklight: Scalable defense for neural networks against query-based black-box attacks. In *USENIX*, 2022.

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[41] Jun Liu, Leo Yu Zhang, Fengpeng Li, Isao Echizen, and Jiantao Zhou. Low-cost hard-label adversarial attack with theoretical foundations. *arXiv preprint*, 2026. URL: https://arxiv.org/abs/2601.14300, arXiv:2601.14300.

[42] Chen Ma, Xiangyu Guo, Li Chen, Jun-Hai Yong, and Yisen Wang. Finding optimal tangent points for reducing distortions of hard-label attacks. *NeurIPS*, 2021.

[43] Thibault Maho, Teddy Furon, and Erwan Le Merrer. Surfree: a fast surrogate-free black-box attack. In *CVPR*, 2021.

[44] Seungyong Moon, Gaon An, and Hyun Oh Song. Parsimonious black-box adversarial attacks via efficient combinatorial optimization. In *ICLR*, 2019.

[45] PyTorch. https://github.com/pytorch/vision, 2023.

[46] Zeyu Qin, Yanbo Fan, Hongyuan Zha, and Baoyuan Wu. Random noise defense against query-based black-box attacks. *Advances in Neural Information Processing Systems*, 34:7650–7663, 2021.

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICLR*, 2021.

[48] Md Farhamdur Reza, Ali Rahmati, Tianfu Wu, and Huaiyu Dai. Cgba: Curvature-aware geometric black-box attack. In *ICCV*, pages 124–133, 2023.

[49] Meng Shen, Changyue Li, Qi Li, Hao Lu, Liehuang Zhu, and Ke Xu. Transferability of white-box perturbations: query-efficient adversarial attacks against commercial dnn services. In *USENIX*, 2024.

[50] Tencent. Tencent general image label. https://cloud.tencent.com/document/product/865/75196, 2024.

[51] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *ICLR*, 2018.

[52] J.H. Van Vleck and D. Middleton. The spectrum of clipped noise. *Proceedings of the IEEE*, 54(1):2–19, 1966.

[53] Viet Quoc Vo, Ehsan Abbasnejad, and Damith C Ranasinghe. Ramboattack: A robust query efficient deep neural network decision exploit. In *NDSS*, 2022.

[54] Gregory K Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991.

[55] Jie Wan, Jianhao Fu, Lijin Wang, and Ziqi Yang. Bounceattack: A query-efficient decision-based adversarial attack by bouncing into the wild. In *IEEE S&P*, 2024.

[56] Feiyang Wang, Xingquan Zuo, Hai Huang, and Gang Chen. Adba: Approximation decision boundary approach for black-box adversarial attacks. *AAAI*, 2025.

[57] Feiyang Wang, Xingquan Zuo, Hai Huang, and Gang Chen. Ttba: Two-third bridge approach for decision-based adversarial attack. In *ICML*, 2025.

[58] Binyan Xu, Xilin Dai, Di Tang, and Kehuan Zhang. One surrogate to fool them all: Universal, transferable, and targeted adversarial attacks with clip. In *ACM CCS*, pages 3087–3101, 2025.

[59] Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *ISBI*, pages 191–195, 2021.

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

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

[62] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *NeurIPS*, 32, 2019.

[63] H. Zhang, Y. Avrithis, T. Furon, and L. Amsaleg. Walking on the edge: Fast, low-distortion adversarial examples. *IEEE TIFS*, 2021.

[64] Baolin Zheng, Peipei Jiang, Qian Wang, Qi Li, Chao Shen, Cong Wang, Yunjie Ge, Qingyang Teng, and Shenyi Zhang. Black-box adversarial attacks on commercial speech platforms with minimal information. In *ACM CCS*, pages 86–107, 2021.

[65] Ziqi Zhou, Yufei Song, Minghui Li, Shengshan Hu, Xianlong Wang, Leo Yu Zhang, Dezhong Yao, and Hai Jin. Darksam: Fooling segment anything model to segment nothing. *NeurIPS*, 37:49859–49880, 2024.

A Detailed Experimental Settings

Victim Models. For CIFAR-10, we select standard ResNet-18 and VGG-16-BN models [45], alongside an adversarially trained WideResNet [14]. For ImageNet-1K, we target a diverse array of high-performing architectures from the Torchvision library [45], including ResNet-50, ViT-B-32, DenseNet-121, EfficientNet-B0, ConvNeXt-Base, Inception-v3, SwinV2-T, DeiT-B, and RegNet-X-8GF. Additionally, we evaluate an adversarially trained WideResNet-50 (trained with $\ell_\infty = 8/255$ [20]). Regarding foundation models [47], we employ the ViT-H-14-CLIPA-336 [31], chosen for its superior zero-shot performance on ObjectNet. For downstream tasks,

Table 10: The untargeted attack performance on CIFAR-10 (VGG-16-BN) under $\ell_\infty = 0.05$ constraint.

Metrics	ASR								Avg.Q								Med.Q							
Max$Q$	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**
5	0.0	0.0	0.5	0.0	1.5	0.0	**4.0**	0.0	-	-	5	-	5	-	5	-	-	-	5	-	5	-	5	-
10	0.0	0.0	2.5	0.0	1.5	0.0	**14.5**	2.5	-	-	7	-	10	-	6	9	-	-	7	-	10	-	6	6
20	2.5	2.5	2.5	0.0	3.5	2.0	14.5	**18.5**	14	14	7	-	19	17	6	14	14	14	7	-	20	17	6	14
50	24.5	24.5	14.0	0.0	8.0	13.0	**30.0**	**30.0**	44	44	33	-	39	35	25	24	48	48	40	-	47	37	24	18
80	32.0	33.0	25.0	4.0	11.0	24.5	**45.5**	41.0	45	47	49	71	49	47	39	35	49	50	41	71	49	49	37	29
100	32.5	37.5	22.5	5.5	18.0	31.5	49.5	**50.5**	45	47	44	73	69	57	43	45	49	50	41		68	55	40	39
150	41.5	40.5	16.0	8.5	26.5	51.0	62.0	**64.0**	82	81	64	91	93	83	61	61	89	89	79	72	98	85	52	56
200	44.0	41.5	27.0	9.0	37.0	66.0	**76.0**	73.0	81	84	115	106	121	104	81	74	89	89	82	107	122	104	66	65
250	45.0	45.0	27.0	15.5	45.5	75.0	80.0	**80.5**	92	85	99	147	143	118	88	88	89	90	80	141	150	117	69	74
300	45.5	45.0	28.0	21.0	57.0	79.0	82.5	**85.5**	89	85	106	185	173	125	94	99	90	90	80	214	179	120	72	82
400	42.0	42.0	29.0	32.5	71.5	86.0	87.5	**90.5**	87	84	120	259	212	144	109	112	89	89	82	298	209	127	79	87
500	44.0	43.0	29.5	40.0	76.5	91.0	**92.5**	91.5	83	86	115	293	227	161	127	116	89	90	81	299	217	132	83	87
1k	47.0	49.5	35.0	64.0	91.0	98.0	**98.5**	**98.5**	137	133	239	491	305	205	163	161	140	139	133	487	265	147	96	92
2k	48.5	48.0	37.0	81.5	98.0	99.5	**100.0**	99.0	201	130	320	657	387	220	178	166	140	139	132	591	288	147	107	94
3k	52.0	47.5	40.0	92.5	99.5	**100.0**	**100.0**	**100.0**	180	136	346	873	393	229	171	188	139	139	134	698	290	147	104	98

Table 11: The untargeted attack performance on ImageNet (ViT-B-32) under $\ell_\infty = 0.05$ constraint.

Metrics	ASR								Avg.Q								Med.Q							
Max$Q$	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	HSJA	BouR	BouT	TtBA	HRayS	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**
5	0.0	0.0	3.5	0.0	0.5	0.0	**8.0**	-	-	-	5	-	5	-	5	-	-	-	5		**5**	-	5	-
10	0.0	0.0	9.0	0.0	0.5	0.0	**18.5**	-	-	-	6	-	10	-	6	-	-	-	6	-	10	-	6	-
20	2.0	2.0	9.0	0.0	0.5	2.0	**19.0**	17.0	14	14	6	-	20	18	12	18	14	14	6	-	20	18	10	19
50	2.5	2.5	12.5	0.0	10.5	17.0	**30.0**	**30.0**	14	14	8	-	42	36	20	25	14	14	6	-	40	36	15	20
80	9.5	9.5	13.0	6.5	17.5	30.0	41.5	**43.5**	48	48	9	75	57	49	32	38	60	60	6	75	57	48	26	27
100	10.0	8.0	15.5	9.0	22.0	35.5	44.5	**48.0**	45	43	25	80	66	55	36	45	60	60	6	80	72	52	29	35
200	11.0	12.0	15.5	15.0	36.5	52.5	64.5	**65.0**	77	82	46	115	103	86	71	71	100	100	6	120	98	74	54	55
300	14.0	11.5	13.0	19.5	44.0	63.5	74.5	**75.5**	85	81	35	143	129	113	97	96	100	99	6	146	117	88	65	71
400	12.5	11.5	14.0	21.0	51.0	70.5	**82.0**	78.5	83	81	85	167	163	136	121	106	99	99	7	149	136	100	80	81
500	11.5	10.0	13.0	25.5	56.5	75.5	**85.5**	84.5	125	70	71	201	193	156	133	130	100	100	7	152	153	116	87	93
1k	14.0	14.5	15.5	41.0	71.0	85.5	89.5	**92.5**	160	124	119	435	299	217	160	176	150	150	13	415	194	146	104	99
3k	15.0	14.0	14.5	67.0	92.0	97.5	98.0	**98.5**	184	132	151	927	602	405	295	272	150	150	139	730	366	178	119	103
5k	16.0	12.0	17.0	76.0	95.5	99.0	**99.5**	**99.5**	340	128	158	1,363	700	453	341	305	151	150	7	972	392	181	119	105
10k	17.5	13.0	14.0	83.0	99.0	**100.0**	**100.0**	**100.0**	1,813	141	153	1,801	961	502	346	349	151	150	72	974	402	186	121	106

Table 12: The untargeted attack performance on ImageNet across diverse architectures under $\ell_\infty = 0.05$ constraint.

Models	EfficientNet-B0			ConvNeXt-base			Inception-v3			SwinV2-T			DeiT-B			RegNet-X-8GF		
Max$Q$	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**
50	24.5	36.5	**44.5**	4.5	17.0	**18.5**	14.5	44.0	**48.5**	10.5	16.5	**18.5**	4.5	8.0	**18.5**	20.5	35.5	**65.0**
500	85.5	**87.0**	85.0	72.5	**75.5**	68.0	85.0	85.5	**87.0**	52.0	**67.5**	**67.5**	49.5	55.0	**62.0**	88.0	**92.0**	91.0

Table 13: The untargeted attack performance comparison on ImageNet under $\ell_2 = 5$ constraint.

Victims	Metrics	ASR								Avg.Q								Med.Q							
	Max$Q$	HSJA	BouR	BouT	SurFree	Tangent	CGBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	HSJA	BouR	BouT	SurFree	Tangent	CGBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	HSJA	BouR	BouT	SurFree	Tangent	CGBA	**Ours$_{opt}$**	**Ours$_{dyn}$**
ResNet-50	100	5.5	5.5	5.5	8.5	8.0	9.0	14.5	**18.0**	39	31	40	43	62	73	39	44	50	19	37	37	44	73	36	39
	500	8.0	6.5	6.5	18.0	18.0	14.0	**34.5**	32.0	104	80	133	168	199	148	171	160	89	79	76	110	150	81	125	80
	1k	9.5	7.5	8.0	26.0	25.0	18.0	**44.5**	**44.5**	162	124	163	351	415	308	295	313	140	136	126	280	310	137	202	243
	5k	10.0	9.0	9.0	52.0	61.5	34.0	**67.5**	63.5	884	525	423	1,410	1,347	1,590	1,088	953	141	138	128	1,062	1,304	958	572	477
Dense-121	100	5.0	5.0	4.0	7.5	7.0	8.5	**15.5**	15.0	42	49	19	46	59	74	32	32	50	47	9	39	45	74	24	22
	500	8.0	6.0	6.5	18.0	13.0	14.0	25.0	**32.5**	113	139	129	306	490	181	132	157	90	88	78	119	150	97	59	123
	1k	5.5	6.5	8.0	22.5	22.0	19.5	33.5	**44.0**	131	113	249	306	490	364	264	304	140	137	131	229	485	291	143	237
	5k	9.5	8.5	7.0	53.0	66.0	37.0	**67.5**	66.5	568	509	367	1,727	1,889	1,469	1,323	1,045	142	138	128	1,334	1,541	1,110	1,052	567
ViT-B-32	100	3.5	6.0	5	5.5	6.5	7.5	15.5	**18.5**	21	43	59	53	64	73	42	45	18	51	39	48	70	73	40	41
	500	10.0	13.5	11.0	13.5	16.5	20.0	**30.5**	**30.5**	157	236	158	171	265	196	138	118	91	182	79	106	219	214	99	76
	1k	11.5	20.5	20.5	19.5	22.0	27.0	36.5	**38.5**	263	486	504	335	493	362	231	247	141	500	490	287	482	247	138	119
	5k	29.5	54.0	57.5	47.0	55.5	52.5	59.5	**62.0**	1,768	1,907	1,985	1,847	1,708	1,331	1,039	1,174	1,296	1,832	1,823	1,422	1,199	871	439	550

we evaluate FasterRCNN-ResNet50-FPN-V2 [45] (object detection) and SAM-ViT-B [35] (segmentation). Finally, we adopt the top-performance NoisyMix [21] and HMany [27] models from RobustBench library [14] for ImageNet-C, and a standard MLP from the MedMNIST library [59, 61] for PathMNIST.

Image Sizes. All inputs are processed as three-channel RGB images. Grayscale images are converted to RGB via channel replication before processing. We use $32 \times 32 \times 3$ for CIFAR-10, $224 \times 224 \times 3$ for ImageNet-1K and ImageNet-C, $336 \times 336 \times 3$ for ObjectNet/CLIP, and $28 \times 28 \times 3$ for PathMNIST. For MS-COCO object detection, we keep the original image resolution, whose height and width are typically around 300–700 pixels. When the image dimensions are not divisible by the effective block size required by our block-wise initialization, we pad the bottom and right borders to the nearest divisible size using replication padding. For SA-1B/SAM, following [65], all images are resized to

Table 14: The ablation study on $\mathbf{d_n}$, $\phi(\hat{\mathbf{d}}_\mathbf{r})$ and PDO. Ours_{opt} serves as the baseline ASR; Increased: **green** background; Decreased: **gray** background.

Datasets	ImageNet (ViT-B-32)				CIFAR-10 (VGG-16-BN)				
Max$Q$	$\textbf{Ours}_{\textbf{opt}}$	w/ $\mathbf{d_n}$	w/ $\phi(\hat{\mathbf{d}}_\mathbf{r})$	w/o P.	Max$Q$	$\textbf{Ours}_{\textbf{opt}}$	w/ $\mathbf{d_n}$	w/ $\phi(\hat{\mathbf{d}}_\mathbf{r})$	w/o P.
5	8.0	0.0	0.5	0.0	5	4.0	0.0	-4.0	0.0
10	18.5	0.0	-4.5	0.0	10	14.5	0.0	-10.5	0.0
50	30.0	-0.5	-5.0	-0.5	50	30.0	-2.5	-13.0	-4.5
80	41.5	-0.5	-6.0	-2.5	80	45.5	-3.5	-14.0	-1.5
100	44.5	4.0	-4.5	0.0	100	49.5	0.0	-12.5	2.5
200	64.5	4.0	-5.5	-6.5	150	62.0	-2.0	-13.0	0.0
300	74.5	1.5	-5.0	-11.0	200	76.0	-3.0	-16.0	-4.0
400	82.0	-1.0	-6.0	-10.0	250	80.0	-1.5	-13.0	-2.5
500	85.5	0.0	-7.5	-8.5	300	82.5	-0.5	-7.0	-1.5
1k	89.5	3.0	0.0	-5.0	400	87.5	-1.0	-7.0	1.0
3k	98.0	0.0	-1.0	-3.5	500	92.5	-3.5	-8.0	-3.0
5k	99.5	0.0	-0.5	-2.5	1k	98.5	-2.5	-5.0	-2.5
10k	100.0	0.0	-0.5	-0.5	2k	100.0	-1.5	-1.0	-1.5

Table 15: The untargeted attack performance on dense prediction tasks. The query limit is 1,000.

Methods	Object Detection			Segmentation		
	ASR	Avg.Q	Med.Q	ASR	Avg.Q	Med.Q
ADBA	39.5	146	73	43.5	364	298
Ours $\mathbf{d_n}$	**70.5**	**65**	27	**68.0**	227	109
Ours_{opt}	65.0	167	36	63.5	161	**36**
Ours_{dyn}	69.5	160	**26**	64.0	**160**	41

$1500 \times 2250 \times 3$ before attack.

Implementation Details. In DPAttack, for low-frequency filtering, the decomposition level is set to its maximum feasible depth, $dl = \lfloor \log_2(\bar{w}) \rfloor$. Empirical analysis shows that while results are marginally sensitive to dl within its valid range, larger values generally yield superior performance. For baselines requiring Monte Carlo gradient estimation (HSJA, Bounce, and Tangent), we tune their batch sizes via grid search to ensure fairness in low-query regimes. This adjustment is necessary because their default batch size (100) leads to 0% ASR whenever the query budget $Q_{\max} < 100$.

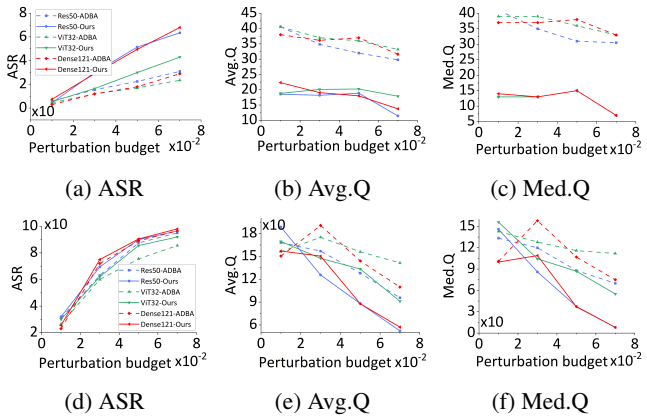
Dense Prediction Tasks. For the results in Tables 8 and 15, we define task-specific success criteria: (i) Object Detection: an attack succeeds if the per-image Average Precision (AP) drops below 0.2; (ii) Segmentation: success is defined as reducing the prediction Intersection over Union (IoU) to < 0.5. Strictly adhering to the hard-label setting, our algorithm only accesses predicted bounding boxes (object detection) or masks (segmentation) during the attack process. Confidence scores are solely utilized for final AP calculation during the evaluation phase and are never used to guide the attack generation.

B The contribution of Directions $\mathbf{d_n}$ and $\mathbf{d_a}$

To further clarify the individual contributions of $\mathbf{d_n}$ and $\mathbf{d_a} = \phi(\hat{\mathbf{d}}_\mathbf{r})$—both designed to improve initial direction in DDM—we introduce two new attack variants that apply only one of them during initialization, keeping all other settings unchanged. These variants, denoted as w/ $\mathbf{d_n}$ and w/ $\mathbf{d_a}$, are evaluated in terms of ASR, as shown in Table 14. We perform the ablation study using Ours_{opt} (i.e. fixed optimal block size with DDM) as the baseline. Query counts are omitted due to their comparable performance. We can observe that, using only $\mathbf{d_n}$ without $\mathbf{d_a}$ even leads to an improvement in ASR for the ViT-B-32. This may be attributed to the high compatibility between the frequency-domain sensitivity of the ViT architecture and the frequency-based sampling strategy of $\mathbf{d_n}$. In contrast, for lower-resolution datasets such as CIFAR-10, incorporating $\mathbf{d_a}$ is crucial for improving the ASR. But when using only $\mathbf{d_a}$ without $\mathbf{d_n}$, the ASR drops significantly across models. These observations indicate that $\mathbf{d_n}$ plays a critical role in enhancing the overall ASR, while $\mathbf{d_a}$ which focuses on low-frequency components, serves to balance performance across diverse black-box models, without substantially compromising the ASR on models where $\mathbf{d_n}$ alone already performs well.

Table 16: The attack performance on ImageNet (adversarially trained WideResNet-50) under different perturbations.

Metrics	ASR			Avg.Q			Med.Q		
ε (ℓ_∞)	ADBA	$\textbf{Ours}_{\textbf{opt}}$	$\textbf{Ours}_{\textbf{dyn}}$	ADBA	$\textbf{Ours}_{\textbf{opt}}$	$\textbf{Ours}_{\textbf{dyn}}$	ADBA	$\textbf{Ours}_{\textbf{opt}}$	$\textbf{Ours}_{\textbf{dyn}}$
0.03	26.0	**28.0**	27.0	158	153	162	56	92	110
0.05	45.0	**50.5**	48.5	143	143	191	58	73	90
0.08	73.0	**73.5**	73.0	96	116	121	41	40	48
0.10	85.5	**87.0**	86.5	74	89	82	35	26	33



(a) ASR (b) Avg.Q (c) Med.Q

(d) ASR (e) Avg.Q (f) Med.Q

Figure 9: Attack performance on ImageNet under different perturbation budgets. The first row shows Max.Q=50, and the second row shows Max.Q=500.

C Additional Results under ℓ_∞ Constraints

C.1 Diverse Perturbation Budgets

Fig. 9 illustrates the attack performance of DPAttack versus ADBA under varying ℓ_∞ constraints. In the low-query regime

(Figs. 9(a), (c), (e)): our method maintains a clear ASR advantage that scales with the perturbation magnitude, while requiring significantly fewer queries to find successful adversarial examples. In the high-query regime (Figs. 9(b), (d), (f)): DPAttack remains superior across different perturbation levels. The only marginal exception occurs at an extremely small bound of $\ell_\infty = 0.01$, where our method shows a slightly higher average query count. This is primarily because DPAttack successfully converts "hard samples" that ADBA fails to attack; these more challenging cases naturally demand higher query counts, which increases the overall average but results in a superior final ASR.

C.2 Diverse Architectures and Domains

To complement the results in Sec. 5.2, we provide extensive evaluations across diverse architectures and datasets. Tables 10 and 11 report the performance on VGG-16-BN (CIFAR-10), as well as ViT-B-32 (ImageNet) across multiple query budgets. More results on ResNet-18 (CIFAR-10) and DenseNet-121 (ImageNet) are presented in our arXiv version [41]. Table 12 further extends these evaluations to a wider range of SOTA ImageNet architectures. Beyond standard natural images, we evaluate DPAttack on corrupted (ImageNet-C) and biomedical (PathMNIST) domains (see Table 17). Our method consistently outperforms SOTA hard-label attacks in both ASR and query efficiency across all tested models, confirming the generalization of our proposed method.

C.3 Robustness against Adversarial Training

Table 16 presents comparative results against an adversarially trained WideResNet-50 across various perturbation levels. Our method achieves a consistently higher ASR than the second-best baseline, ADBA, while maintaining competitive query efficiency. Notably, the ASR gains—though numerically modest—represent a significant improvement in the challenging context of attacking robust models.

C.4 Generalization on Dense Prediction Tasks

Beyond the results in Table 8 of the main manuscript, which demonstrate the superiority of our method under a limited query budget (50), Table 15 further shows that our approach maintains a performance margin of over 30% even when the query limit is relaxed to 1,000.

D Additional Results Under ℓ_2 Constraint

While DPAttack is primarily tailored for the structural characteristics of ℓ_∞ perturbations, we evaluate its performance under ℓ_2 constraints to demonstrate its versatility. Table 13 compares DPAttack against SOTA methods specifically optimized for the ℓ_2 metric, such as SurFree and Tangent. Despite its ℓ_∞-focused design, DPAttack achieves superior ASR and query efficiency—particularly in low-query regimes—validating its robust generalization across different distance metrics.

Table 17: The untargeted attack performance on ImageNet-C and PathMNIST, respectively. Max.Q= 50.

Datasets	Models	ASR			Avg.Q			Med.Q		
		ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**	ADBA	**Ours$_{opt}$**	**Ours$_{dyn}$**
ImgNC	NMix	33.5	**50.0**	48.5	30	**13**	22	29	**6**	20
	HMany	43.5	**59.0**	56.5	31	**16**	22	29	**6**	20
PMNIST	MLP	21.0	**35.0**	29.5	37	**31**	35	37	**31**	36

Table 18: Comparison with CA [29] under the same settings on ImageNet with ResNet-18. The values are CA/**Ours$_{dyn}$**.

Defense	ASR ↑	Det$R$ ↓	Avg.Q↓	Med.Q↓	ℓ_2 ↓
Blacklight	100% / 100%	0% / 0%	603 / **16**	603 / **9**	33.0 / **31.3**
RAND-Pre	100% / 100%	–	603 / **29**	603 / **16**	32.2 / **28.2**
RAND-Post	100% / 100%	–	603 / **20**	603 / **16**	32.7 / **29.2**

E Comparison with Certifiable Attack (CA)

We further compare DPAttack with the recent Certifiable Attack (CA) [29], which represents a different attack paradigm based on randomized adversarial distributions. Unlike empirical hard-label attacks that search for a single adversarial example, CA constructs an adversarial distribution with a guaranteed attack success probability, and adversarial examples sampled from this distribution can be used without final verification queries. This inherent randomness makes CA particularly effective against stateful detection mechanisms such as Blacklight.

CA provides two localization strategies: Smoothed Self-Supervised Perturbation (SSSP) localization and binary-search localization. SSSP leverages a pretrained feature extractor as an additional surrogate prior, while binary-search localization does not rely on such pretrained adversarial or feature priors. Since DPAttack also does not use adversarial-example priors or surrogate feature extractors, we compare against the binary-search version of CA for a closer surrogate-free comparison.

Table 18 reports the comparison under the same settings as CA on ImageNet with ResNet-18. DPAttack achieves the same ASR and detection rate as CA across Blacklight, RAND-Pre [46], and the RAND-Post setting used in CA, while requiring about 25× fewer queries and smaller perturbations. Notably, CA is evaluated under a more relaxed unrestricted perturbation budget. These results suggest that although CA provides a strong certifiable and verification-free attack paradigm, DPAttack offers a more query-efficient empirical alternative when the goal is to obtain successful hard-label AEs under strict query budgets.