QUERY EFFICIENT BLACK-BOX ADVERSARIAL ATTACK VIA DATASET DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Score-based black-box attacks seek effective adversarial perturbations under limited access to the victim model, making query efficiency a central challenge. Existing methods often improve query efficiency by incorporating prior information, such as structured search priors or surrogate models to guide the search toward promising perturbation directions. However, the effectiveness of such guidance depends strongly on the quality and availability of the underlying prior. To provide a richer and alternative source of guidance, we propose **DDAttack**, which leverages dataset distillation (DD) as prior information for gradient-free adversarial search. DD condenses class-discriminative information from large datasets into a compact set of synthetic examples, providing informative input-space cues that can guide adversarial optimization. DDAttack exploits this distilled information through two complementary stages: *distillation-guided initialization*, which constructs a promising starting perturbation from distilled examples, and *distillation-guided search*, which aggregates promising candidate perturbations to guide subsequent updates. Extensive experiments on adversarially trained models demonstrate that distilled examples provide effective prior information for score-based black-box attacks. The source code is included in the supplementary material and will be publicly available upon acceptance.

1 INTRODUCTION

Deep neural networks (DNNs) are well known to be susceptible to adversarial attacks (Szegedy et al., 2013). Depending on the attacker's knowledge of the victim model, adversarial attacks are commonly categorized as *white-box attacks* (Madry et al., 2017; Goodfellow et al., 2014) or *black-box attacks* (Zheng et al., 2025; Andriushchenko et al., 2020). White-box attacks directly exploit model parameters and gradients, whereas black-box attacks must construct adversarial perturbations without access to such internal information. Black-box attacks are particularly relevant when models are accessible only through prediction interfaces and provide a complementary means of robustness evaluation when model gradients are unavailable or unreliable.

Among different black-box attack paradigms, score-based attacks (Andriushchenko et al., 2020) assume access to the victim model's output information, such as logit distribution, and use this limited feedback to guide the search for adversarial perturbations. A central challenge is therefore how to extract informative search directions from a limited number of victim queries. Existing methods address this challenge through gradient estimation (Ilyas et al., 2018), structured random search (Andriushchenko et al., 2020), and auxiliary priors obtained from surrogate models (Lord et al., 2022; Cheng et al., 2019). These approaches highlight the importance of informative priors in gradient-free adversarial optimization, and motivate the exploration of richer sources of prior information that can provide more effective guidance for gradient-free adversarial search.

Dataset distillation (DD) was originally introduced to condense the training-relevant knowledge of a large dataset into a compact set of synthetic examples that can effectively support DNN training (Wang et al., 2018). More recent studies have shown that the information encoded in distilled data can benefit a broader range of learning scenarios beyond standard model training, including federated learning, neural architecture search, and continual learning (Lei & Tao, 2023; Yu et al., 2024). These findings suggest that distilled examples capture rich and transferable information about the underlying data distribution and task. However, their potential for adversarial attacks, particu-

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

larly score-based black-box attacks, remains largely unexplored. We argue that the rich information contained in distilled examples can provide useful guidance for adversarial search, offering a new source of prior knowledge.

Motivated by the fact that distilled examples compactly encode class-discriminative information in the input space, we propose **DDAttack**, a score-based attack that leverages this information as a prior for adversarial search. Intuitively, distilled examples from competing classes provide class-relevant directions that can help guide the attack toward alternative decision regions. DDAttack incorporates distilled information in two complementary stages: *distillation-guided initialization* and *distillation-guided search*. During initialization, DDAttack leverages the prior in the distilled samples to obtain a stronger initial point for the subsequent adversarial search. During distillation-guided search, DDAttack aggregates the most promising candidate perturbations into a guidance direction and uses it to guide the local updates more effectively. We evaluate DDAttack on ImageNet-1K against a diverse set of adversarially robust classifiers. Under fixed query budgets, DDAttack consistently outperforms existing representative score-based black-box attacks across different victim architectures and perturbation budgets, showing that the information encoded in distilled examples can serve as an effective prior for improving query-efficient adversarial search.

Our main contributions are summarized as follows:

- We investigate dataset distillation as a new source of prior knowledge for score-based black-box adversarial attacks, leveraging the information encoded in distilled samples to guide adversarial search.
- We propose **DDAttack**, a strategy that leverages distilled priors to improve adversarial search efficiency under limited query budgets.
- Comprehensive experiments show that DDAttack consistently outperforms existing representative score-based attacks across different victim architectures and perturbation budgets.

2 Related Work

Adversarial attack. Adversarial attacks are commonly studied under white-box and black-box settings. White-box methods, such as FGSM (Goodfellow et al., 2014), PGD (Madry et al., 2017), and APGD (Croce & Hein, 2020), directly exploit model gradients and typically assume full access to the victim model's parameters and gradients. However, this assumption can limit their applicability in practical scenarios where such internal information is unavailable. In contrast, black-box methods (Zheng et al., 2025; Cheng et al., 2024; Yin et al., 2023a; Feng et al., 2022; Cheng et al., 2021) construct adversarial examples without access to the victim's parameters or gradients, and are commonly divided into score-based attacks, which query the model's output scores and decision-based attacks, which observe only the predicted label.

Score-based attacks have received considerable attention because output scores provide richer feedback than hard-label predictions, enabling more informative adversarial search while retaining black-box access. NES (Ilyas et al., 2018) estimates attack gradients using natural evolution strategies with randomly sampled perturbations. Bandit (Ilyas et al., 2019) incorporates temporal and spatial priors into a bandit optimization framework to improve query efficiency. ZOsign-SGD (Liu et al., 2019) updates perturbations using the signs of zeroth-order gradient estimates, while Sign-Hunter (Al-Dujaili & O'Reilly, 2020) directly estimates gradient signs through adaptive search over sign configurations. Square Attack (Andriushchenko et al., 2020) adopts a structured random-search strategy based on localized square-shaped perturbation updates. To improve query efficiency, several methods further leverage auxiliary surrogate models as transferable priors to guide adversarial search (Cheng et al., 2019; Guo et al., 2019; Yang et al., 2020; Huang & Zhang, 2020; Lord et al., 2022). In the decision-based setting, ADBA-md (Wang et al., 2025) relies only on predicted labels and compares perturbation directions through median-based approximate decision boundaries. Different from these approaches, DDAttack explores dataset distillation as a source of class-related prior information for guiding score-based adversarial search without relying on an auxiliary surrogate model.

Dataset distillation. Dataset distillation aims to condense the training-relevant information of a large dataset into a compact set of synthetic examples that can support effective model training (Wang et al., 2018; Kim et al., 2022; Cazenavette et al., 2023; Du et al., 2023). Subsequent

approaches construct distilled datasets by aligning model training trajectories (Cazenavette et al., 2022), matching gradients between real and synthetic data (Zhao et al., 2020), or matching feature distributions (Zhao & Bilen, 2023). To extend dataset distillation to larger-scale settings, SRe2L (Yin et al., 2023b) decouples the condensation process into squeezing, recovery, and relabeling stages, enabling ImageNet-scale distillation. RDED (Sun et al., 2024) further emphasizes the diversity and realism of distilled examples to improve large-scale dataset distillation. While these methods primarily optimize distilled data for model training, our work explores a different use of distilled examples: serving as image-space priors to guide black-box adversarial search.

3 Proposed method

3.1 Preliminaries

Problem setting. Let $f : \mathbb{X}^{C\times H\times W} \to \mathbb{R}^K$ denote the victim classifier, where C, H, and W represent the input dimensions and K is the number of classes. For input $x \in \mathbb{X}^{C\times H\times W}$, we define the victim prediction as

$$\hat{y}(x) = \arg\max_{k\in 1,\dots,K} f_k(x). \tag{1}$$

Given an input x with ground-truth label y, we define the adversarial example as

$$\widetilde{x}(\delta) = \mathrm{clip}_{[0,1]}(x+\delta), \tag{2}$$

where δ is the adversarial perturbation, and the clip operation ensures the perturbed input remains in the pixel range. An untargeted adversarial attack seeks a perturbation δ such that

$$\hat{y}(\widetilde{x}(\delta)) \neq y, \qquad \|\delta\|_\infty \leq \epsilon, \tag{3}$$

where ϵ is the perturbation budget. In the score-based black-box setting, the attacker can query the victim's output scores but has no access to its parameters or gradients. The attack must therefore identify an effective perturbation within an attack-query budget Q.

Dataset distillation samples. Throughout the attack, DDAttack assumes access to a distilled dataset $\mathcal{D}_{\mathrm{DD}}$:

$$\mathcal{D}_{\mathrm{DD}} = \left\{ \{s_{k,j}\}_{j=1}^{N_k} \right\}_{k=1}^{K}, \tag{4}$$

where $s_{k,j}$ is the j-th of N_k distilled samples of class k. $\mathcal{D}_{\mathrm{DD}}$ is precomputed from the ImageNet-1K training set by dataset distillation methods before the attack.

3.2 Distillation-guided initialization

The key idea of DDAttack is to leverage the rich class-related information encoded in distilled samples as a prior to improve the search efficiency of DDAttack queries. We incorporate this guidance into two phases: **distillation-guided initialization** and **distillation-guided search**. The initialization stage aims to construct a promising initial perturbation δ_0, providing a stronger starting point for the subsequent adversarial search.

Selecting distilled references. Naively, DDAttack could leverage distilled samples from all $K-1$ non-ground-truth classes to construct candidate initializations. However, as the number of classes increases, this strategy would yield substantially more candidates, and thus require more queries to the victim model. To improve query efficiency, DDAttack restricts distilled references to the L most competitive non-ground-truth classes and samples only N_{ref} examples from each class. For an input x with label y, we query the victim model to obtain $f(x)$, select the L highest-scoring non-target classes as $\mathcal{C}(x)$, and randomly sample N_{ref} distilled examples from each selected class:

$$\mathcal{S}(x) = \left\{ \{s_{k,j}\}_{j=1}^{N_{\mathrm{ref}}} \right\}_{k\in\mathcal{C}(x)}. \tag{5}$$

Then, we use $\mathcal{S}(x)$ to construct candidate perturbations for initialization.

Constructing candidate perturbations. Rather than treating distilled samples as target images to reconstruct, we use them as directional priors that indicate class-related directions in the input space. For each reference $s_i \in \mathcal{S}(x)$, we construct a perturbation that maximizes the cosine similarity of

Algorithm 1 Distillation-guided initialization

Input: Victim f; input x with label y; distilled dataset $\mathcal{D}_{\mathrm{DD}}$; budget ϵ; number of competing classes L; references per class N_{ref}; balance weight λ; steps R; step size α; retained set size G
Output: Guidance set $\mathbf{\Delta}_G$ and initialization (δ_0, m_0), or adversarial example $\widetilde{x}(\delta_i)$

$\mathcal{C}(x) \leftarrow$ the L highest-scoring classes in $f(x)$ excluding y
$\mathcal{S}(x) \leftarrow \big\{\{s_{k,j}\}_{j=1}^{N_{\mathrm{ref}}}\big\}_{k\in\mathcal{C}(x)}$ ▷ Select references, Eq. 5
for each reference $s_i \in \mathcal{S}(x)$ **do**
  $\delta_i \leftarrow 0$
  for $r = 1$ **to** R **do**
    $\delta_i \leftarrow \mathrm{clip}_{[-\epsilon,\epsilon]}\big(\delta_i - \alpha\ \mathrm{sign}(\nabla_{\delta_i}\mathcal{L}(\delta_i))\big)$ ▷ Optimize candidate perturbations, Eq. 7
  end for
end for
for each candidate δ_i **do**
  $m_i \leftarrow f_y(\widetilde{x}(\delta_i)) - \max_{k\neq y} f_k(\widetilde{x}(\delta_i))$ ▷ Evaluate candidate, one query, Eq. 8
  if $m_i < 0$ **then return** $\widetilde{x}(\delta_i)$ ▷ Attack succeeds
end for
$\mathcal{I}_G \leftarrow$ indices of the G smallest margins m_i
$\mathbf{\Delta}_G \leftarrow \{(\delta_i, m_i)\}_{i\in\mathcal{I}_G}$ ▷ Selected guidance set, Eq. 9
$i^\star \leftarrow \arg\min_{i\in\mathcal{I}_G} m_i; \quad (\delta_0, m_0) \leftarrow (\delta_{i^\star}, m_{i^\star})$ ▷ Selected initial perturbation, Eq. 10
return $\mathbf{\Delta}_G,\ (\delta_0, m_0)$

the perturbed input $\widetilde{x}(\delta_i)$ and the reference s_i. Meanwhile, to encourage the candidate to effectively exploit the available perturbation budget, we also reduce its similarity to the original input x. We therefore define the objective

$$\mathcal{L}(\delta_i) = \lambda\ \cos\big(\widetilde{x}(\delta_i), x\big) - \cos\big(\widetilde{x}(\delta_i), s_i\big), \tag{6}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity between flattened images, and $\lambda \geq 0$ is the balancing hyperparameter. Note that this objective provides an image-space construction rule for converting distilled references into candidate perturbations, without requiring any victim queries.

Starting from $\delta_i^{(0)} = 0$, we minimize $\mathcal{L}$ using R projected sign-gradient steps. At time step $r \in \{1, \ldots, R\}$, the new perturbation $\delta_i^{(r+1)}$ is calculated as

$$\delta_i^{(r+1)} = \mathrm{clip}_{[-\epsilon,\epsilon]}\left(\delta_i^{(r)} - \alpha\ \mathrm{sign}\left(\nabla_{\delta_i}\mathcal{L}\big(\delta_i^{(r)}\big)\right)\right), \tag{7}$$

where α is the step size. Since the gradients come only from the explicit objective in Eq. 6, this construction requires no victim queries, victim gradients, or surrogate model.

Selecting initial perturbation. Since image-space alignment does not directly determine the victim's prediction, we evaluate the constructed candidates using the victim's output logits. For each candidate perturbation δ_i, we compute the untargeted margin

$$m_i = f_y(\tilde{\mathbf{x}}(\delta_i)) - \max_{k\neq y} f_k(\tilde{\mathbf{x}}(\delta_i)). \tag{8}$$

Note that this candidate evaluation phase is the only phase that requires victim queries during the distillation-guided initialization. The query budget equals the size of the reference set $\mathcal{S}(x)$.

During this phase, if any candidate $\widetilde{x}(\delta_i)$ successfully changes the victim model's prediction, the attack terminates. Otherwise, we retain the G perturbation-margin pairs with the smallest margins

$$\mathbf{\Delta}_G = \{(\delta_i, m_i)\}_{i\in\mathcal{I}_G}, \tag{9}$$

where $\mathcal{I}_G$ denotes the index set of the candidates with the smallest values of m_i. These retained pairs are subsequently used to guide the iterative search. Furthermore, among the retained candidates, we select the perturbation with the smallest margin as the initialization

$$i^\star = \arg\min_{i\in\mathcal{I}_G} m_i, \qquad \delta_0 = \delta_{i^\star}, \qquad m_0 = m_{i^\star}. \tag{10}$$

With the best candidate (δ_0, m_0), we initialize the distillation-guided search from $\widetilde{x}(\delta_0)$. The distillation-guided initialization requires a total of $|\mathcal{S}(x)| + 1$ victim queries: one query to obtain the clean prediction and $|\mathcal{S}(x)|$ queries to evaluate the candidate perturbations. We summarize the procedure of distillation-guided initialization in Algo. 1.

Algorithm 2 Distillation-guided search

Input: Victim f; input x with label y; guidance set $\mathbf{\Delta}_G = \{(\delta_i, m_i)\}_{i\in\mathcal{I}_G}$; initialization (δ_0, m_0); budget ϵ; remaining queries $T = Q - |\mathcal{S}(x)| - 1$; patch-fraction bounds $p_{\min}, p_{\max}$; margin threshold θ; temperatures τ_d, τ_p; initial guidance probability γ_0

Output: Adversarial example $\widetilde{x}(\delta)$, or failure after T queries

$w_i \leftarrow \dfrac{e^{-m_i/\tau_d}}{\sum_{j\in\mathcal{I}_G} e^{-m_j/\tau_d}}$ for $i \in \mathcal{I}_G$; $\quad \bar{\delta} \leftarrow \sum_{i\in\mathcal{I}_G} w_i\delta_i$ ▷ Guidance direction, Eq. 11

$p_0 \leftarrow p_{\min} + (p_{\max} - p_{\min})\ \mathrm{sigmoid}\big((m_0 - \theta)/\tau_p\big)$ ▷ Adaptive patch size, Eq. 12

for $t = 0$ **to** $T - 1$ **do** ▷ Iterative search

  $p_t \leftarrow \mathrm{PatchSchedule}(p_0, t)$

  $a_t \leftarrow \min\{\min(H, W),\ \max\{1, \mathrm{round}(\sqrt{p_t HW})\}\}$ ▷ Current patch size, Eq. 13

  Sample a random $a_t \times a_t$ square $\mathcal{R}_t$; build mask P_t ▷ Region proposal, Eq. 14

  $\gamma_t \leftarrow \frac{\gamma_0}{2}\big[1 + \cos\big(\pi t/\max\{1, T-1\}\big)\big]$ ▷ Cosine guidance decay, Eq. 15

  $v_t \leftarrow \mathrm{sign}(\bar{\delta})$ with probability γ_t, else u_t (i.i.d. per-channel signs from $\{-1, +1\}$) ▷ Eq. 16

  $\hat{\delta}_t \leftarrow \mathrm{clip}_{[-\epsilon,\epsilon]}\big(\delta_t + 2\epsilon\, P_t \odot v_t\big)$ ▷ Generated perturbation, Eq. 17

  $\hat{m}_t \leftarrow f_y(\widetilde{x}(\hat{\delta}_t)) - \max_{k\neq y} f_k(\widetilde{x}(\hat{\delta}_t))$ ▷ Evaluate candidate, one query, Eq. 8

  $(\delta_{t+1}, m_{t+1}) \leftarrow (\hat{\delta}_t, \hat{m}_t)$ if $\hat{m}_t < m_t$, else (δ_t, m_t) ▷ Eq. 18

  if $m_{t+1} < 0$ **then return** $\widetilde{x}(\delta_{t+1})$ ▷ Attack succeeds

end for

return FAIL

3.3 DISTILLATION-GUIDED SEARCH

After initialization, DDAttack leverages the distilled prior during iterative search. Given the guidance set $\mathbf{\Delta}_G$, we use the perturbations δ_i to construct a guidance direction. Also, we separately use the initialization margin m_0 to determine the initial search scale. We use stronger guidance in the early stage to improve search efficiency, and gradually shift toward random exploration to avoid over-reliance on directions that may not align well with the victim model.

Guidance direction. The guidance set $\mathbf{\Delta}_G$ provides a set of prior directions. To leverage such prior effectively, for each candidate $(\delta_i, m_i) \in \mathbf{\Delta}_G$, we construct a weighted guidance direction $\bar{\delta}$ with

$$w_i = \frac{e^{-m_i/\tau_d}}{\sum_{j\in\mathcal{I}_G} e^{-m_j/\tau_d}}, \qquad \bar{\delta} = \sum_{i\in\mathcal{I}_G} w_i\delta_i, \tag{11}$$

where τ_d is the temperature hyperparameter and is set to 1.

Region proposal of search. Following Square attack (Andriushchenko et al., 2020), DDAttack uses square patch proposals. While Square Attack uses a fixed initial patch-area fraction $p_{\max}$, DDAttack adapts this fraction according to the initialization margin m_0. Specifically, the initial patch size is

$$p_0 = p_{\min} + (p_{\max} - p_{\min})\,\mathrm{sigmoid}\left(\frac{m_0 - \theta}{\tau_p}\right), \tag{12}$$

where $p_{\min}$ and $p_{\max}$ bound the initial patch size, and τ_p is the temperature hyperparameter set to 0.5. Given an initial adversarial input $\widetilde{x}(\delta)$, a smaller initialization margin m_0 yields a smaller patch size, whereas a larger margin yields a broader exploration.

During the search, we progressively reduce the patch area fraction p_t following the piecewise schedule of Square Attack (Andriushchenko et al., 2020), and the patch side length at iteration t is

$$a_t = \min\Big\{\min(H, W),\ \max\big\{1,\ \mathrm{round}\big(\sqrt{p_t\, HW}\big)\big\}\Big\}, \tag{13}$$

where $H \times W$ is the resolution of the image. We randomly select a square region $\mathcal{R}_t$ of side length a_t and define its binary mask P_t as

$$P_t(u, v) = \begin{cases} 1, & (u, v) \in \mathcal{R}_t, \\ 0, & \text{otherwise}, \end{cases} \qquad P_t \in \{0, 1\}^{H\times W}. \tag{14}$$

More detailed implementation details about the patch-area scheduling are included in the Appendix.

Generation of perturbation proposals. Although $\bar{\delta}$ provides an informative guidance direction, relying on this fixed prior throughout the entire search may restrict perturbation diversity and limit attack effectiveness. Thus, we gradually decrease the probability of following $\bar{\delta}$ as the search proceeds, allowing the attack to transition from distilled guidance toward broader random exploration. Let $T = Q - |\mathcal{S}(x)| - 1$ be the remaining query budget after distillation-guided initialization. At step $t \in \{0, \dots, T-1\}$, the guidance probability follows a cosine decay:

$$\gamma_t = \frac{\gamma_0}{2}\left[1 + \cos\left(\frac{\pi t}{\max\{1, T-1\}}\right)\right], \tag{15}$$

where $\gamma_0 \in [0, 1]$ is the initial guidance probability set to 0.5. At each search step, we choose the proposal direction as

$$v_t = \begin{cases} \operatorname{sign}(\bar{\delta}), & \text{with probability } \gamma_t, \\ u_t, & \text{with probability } 1 - \gamma_t, \end{cases} \tag{16}$$

where u_t is a random sign pattern with one sign from $\{-1, +1\}$ sampled independently per channel. Given the current perturbation δ_t, we propose the new perturbation $\hat{\delta}_t$ as

$$\hat{\delta}_t = \operatorname{clip}_{[-\epsilon,\epsilon]}\left(\delta_t + 2\epsilon\, P_t \odot v_t\right), \tag{17}$$

where P_t is the mask in Eq. 14 and $\odot$ is elementwise multiplication. We query the victim on $\widetilde{x}(\hat{\delta}_t)$ to obtain its margin $\hat{m}_t$ (Eq. 8), and accept the proposal only if it improves the current margin m_t:

$$(\delta_{t+1}, m_{t+1}) = \begin{cases} (\hat{\delta}_t, \hat{m}_t), & \hat{m}_t < m_t, \\ (\delta_t, m_t), & \text{otherwise.} \end{cases} \tag{18}$$

Distillation-guided search ends when an adversarial example is found or the query budget is exhausted. The summary of this procedure is shown in Algo. 2. Moreover, we provide a theoretical grounding for DDAttack's design in Appendix A.

4 Experiments

4.1 Experiment setup

Dataset and victim models. All experiments are conducted on the ImageNet-1K dataset (Deng et al., 2009) at the standard input resolution of 224×224. We evaluate black-box attack methods against a diverse set of robust classifiers covering both isotropic and hierarchical Transformer architectures, including ViT- and Swin-based models trained with **Pretrain + AT** (Singh et al., 2023), **input gradient regularization** (Rodríguez-Muñoz et al., 2024), **ARES-Bench adversarial training** (Liu et al., 2025), and **MIMIR** (Xu et al., 2023).

Evaluated attack methods. We compare DDAttack with representative black-box attacks as introduced in Sec. 2, including **NES** (Ilyas et al., 2018), **Bandit** (Ilyas et al., 2019), **ZOsign-SGD** (Liu et al., 2019), **SignHunter** (Al-Dujaili & O'Reilly, 2020), **Square Attack** (Andriushchenko et al., 2020), and **ADBA-md** (Wang et al., 2025). Additionally, we compare DDAttack with representative surrogate-based attacks, **P-RGF** (Cheng et al., 2019) and **SimBA-ODS** (Tashiro et al., 2020).

Metrics and evaluation details. We report clean accuracy, robust accuracy, and attack success rate (ASR). Clean accuracy is measured with test images without adversarial perturbation, while robust accuracy is the portion of the samples that remain correctly classified after the attack. ASR is the percentage of originally correctly classified samples that are misclassified after the attack. Following the evaluation protocol in Robustbench (Croce et al., 2021), we report the accuracy and ASR on 5,000 ImageNet validation samples.

Implementation details. We conduct untargeted attacks under an ℓ_∞ constraint with $\epsilon \in \{4/255, 8/255\}$ and a maximum query budget of $Q = 5{,}000$ per image. For DDAttack, we use precomputed ImageNet-1K distilled datasets generated by RDED (Sun et al., 2024) for major experiments and the representative large-scale distillation method SRe2L (Yin et al., 2023b) for ablation study in Sec. 4.3. During distillation-guided initialization, we select the $L = 10$ most competitive

Table 1: Robustness evaluation of black-box attacks on ImageNet-1K under ℓ_∞ perturbations with $\epsilon = 8/255$. Following the standard RobustBench protocol (Croce et al., 2021), we evaluate 5,000 validation images, with each attack allowed up to 5,000 queries to the victim model per image. We report clean accuracy and robust accuracy after attack, where lower robust accuracy indicates a stronger attack. The best result for each victim model is shown in bold.

Victim model	Clean Acc.	Robust Acc. ↓						
		NES	Bandit	ZOsign	SignHunter	ADBA-md	Square	DDAttack
ViT-S (Pretrain + AT)	72.58	69.16	55.26	70.56	59.98	58.60	48.92	**47.34**
ViT-B (Pretrain + AT)	76.30	73.14	60.82	74.56	65.22	64.00	55.22	**53.36**
Swin-B (Input Grad Reg.)	77.76	73.82	62.22	75.58	66.18	65.30	56.02	**54.96**
Swin-L (Input Grad Reg.)	79.36	76.50	64.54	77.78	69.00	66.90	59.14	**57.64**
Swin-B (ARES-Bench AT)	76.16	73.16	61.38	74.30	65.26	66.10	56.48	**54.98**
Swin-L (ARES-Bench AT)	78.92	75.74	64.96	77.22	68.74	68.70	60.32	**58.64**
Swin-B (MIMIR)	76.62	73.10	61.60	74.44	65.46	66.80	55.92	**53.72**
Swin-L (MIMIR)	78.62	75.86	64.96	77.12	68.62	68.30	60.22	**58.28**

Table 2: Robustness evaluation under ℓ_∞ perturbations with $\epsilon = 4/255$. We follow the same evaluation protocol as Table 1. Lower robust accuracy indicates a stronger attack, and the best result is shown in bold.

Victim model	Clean Acc.	Robust Acc. ↓						
		NES	Bandit	ZOsign	SignHunter	ADBA-md	Square	DDAttack
ViT-S (Pretrain + AT)	72.58	71.28	64.20	71.96	66.80	67.80	62.80	**60.98**
ViT-B (Pretrain + AT)	76.30	75.38	69.08	75.72	71.16	71.40	67.72	**66.42**
Swin-B (Input Grad Reg.)	77.76	76.34	69.70	76.88	72.20	72.40	68.40	**66.68**
Swin-L (Input Grad Reg.)	79.36	78.46	72.66	78.90	74.90	74.60	71.06	**69.76**
Swin-B (ARES-Bench AT)	76.16	75.00	68.98	75.52	70.98	71.40	67.74	**66.32**
Swin-L (ARES-Bench AT)	78.92	77.68	71.82	78.34	74.12	74.00	70.38	**69.16**
Swin-B (MIMIR)	76.62	75.22	69.06	75.70	71.44	72.70	67.68	**66.50**
Swin-L (MIMIR)	78.62	77.56	72.20	78.04	73.88	73.50	71.00	**69.54**

classes and sample $N_{\text{ref}} = 10$ distilled references from each class, resulting in $|\mathcal{S}(x)| = 100$ candidate perturbations. We optimize the cosine objective with $\lambda = 1$ for $R = 5$ projected sign-gradient steps using a step size of $\alpha = \epsilon$. The candidate with the smallest untargeted margin is used to initialize the search, while the $G = 10$ candidates with the smallest margins form the guidance set. During distillation-guided search, we use $p_{\min} = 0.075$ and $p_{\max} = 0.8$ for adaptive initialization of the patch area. The total query budget includes the 101 victim queries used to evaluate initialization candidates, leaving 4,899 queries for iterative search. More details about the implementation of baseline methods and the distilled dataset are included in the Appendix C.

4.2 Results and analysis

Tables 1 and 2 compare DDAttack with existing black-box attacks under ℓ_∞ perturbation budgets of $\epsilon = 8/255$ and $4/255$, respectively. DDAttack consistently achieves the lowest robust accuracy across all evaluated victim models and both perturbation budgets. This improvement is observed across different Transformer architectures and robustness training strategies, indicating that the effectiveness of distilled guidance is not limited to a particular model architecture or defense.

At $\epsilon = 8/255$, DDAttack reduces robust accuracy by an average of 1.67 percent compared with the strongest competing baseline across the evaluated models. The improvement is consistent for all eight victim models, with reductions ranging from 1.06 to 2.20 percent. A similar trend is observed under the more restrictive $\epsilon = 4/255$ setting. These results show that distilled examples provide useful prior information for guiding adversarial search and improve attack effectiveness under a fixed query budget, even when the allowable perturbation magnitude is reduced.

4.3 Ablation Studies

Effect of distillation-guided initialization and search. We ablate the two components of DDAttack by independently replacing the distillation-guided initialization and search with their random

Table 3: Ablation of the two DD-guided stages under ℓ_∞ with $\epsilon = 8/255$. Note that using both random initialization and random search is equivalent to Square Attack.

Victim model	DDAttack Stages		Robust Acc. ↓
	Initialization	Search	
ViT-B (Pretrain+AT)	Random	Random	55.22
	DD-guided	Random	54.80
	Random	DD-guided	53.88
	DD-guided	DD-guided	53.36
Swin-B (MIMIR)	Random	Random	55.92
	DD-guided	Random	55.34
	Random	DD-guided	54.18
	DD-guided	DD-guided	53.72

Table 4: Ablation on the source of prior information used by DDAttack. We report robust accuracy (%) under an ℓ_∞ perturbation budget of $\epsilon = 8/255$, comparing randomly sampled real training images with distilled samples generated by SRe2L and RDED.

Victim model	Reference Source		
	Real sample	SRe2L	RDED
ViT-B (Pretrain + AT)	54.10	53.52	53.36
Swin-B (Input Grad Reg.)	55.98	55.14	54.96
Swin-B (ARES-Bench AT)	56.24	55.28	54.98
Swin-B (MIMIR)	55.64	54.08	53.72

counterparts. All variants use the same total budget of 5,000 victim queries and the same ℓ_∞ perturbation budget of $\epsilon = 8/255$. When both components are random, the attack reduces to Square Attack, which serves as our baseline. Note that for random initialization and dd-guided search, we generate $\mathbf{\Delta}_G$ with 101 queries using Eq. 9, and use the remaining 4,899 queries for the distillation-guided search, with random Gaussian initialization δ_0.

As shown in Table 3, both components improve attack effectiveness. On ViT-B, distillation-guided initialization alone lowers the robust accuracy from 55.22% to 54.80%, and distillation-guided search alone lowers it to 53.88%. Combining both achieves the lowest robust accuracy of 53.36%. On Swin-B (MIMIR), the full DDAttack similarly reduces the robust accuracy from 55.92% to 53.72%. On both models, the search component contributes more than the initialization, and the combination yields the strongest attack. These results suggest that the two components are complementary: distillation-guided initialization provides a stronger starting point, while distillation-guided search keeps exploiting the distilled prior to improve perturbation updates throughout the search.

Effect of distilled references. Table 4 compares different sources of prior information used by DDAttack, including randomly sampled real training images and distilled samples generated by SRe2L (Yin et al., 2023b) and RDED (Sun et al., 2024). SRe2L is a representative large-scale dataset distillation method designed for ImageNet-scale condensation, while RDED is a more recent strong large-scale dataset distillation method that further improves the quality of distilled samples. Across all evaluated victim models, both distilled datasets consistently outperform the real-sample baseline, indicating that the advantage of DDAttack does not simply come from using class-related reference images. Among the two distillation methods, RDED consistently achieves the lowest robust accuracy, while SRe2L also provides clear improvements over real samples. These results suggest that distilled samples provide more effective guidance for adversarial search than randomly selected real images, and that higher-quality distilled data can further improve attack effectiveness. Meanwhile, the relatively small gap between SRe2L and RDED indicates that DDAttack is not strongly dependent on a particular distillation method.

4.4 Discussion

Query efficiency. Fig. 1 compares the evolution of ASR as the query budget increases on ViT-B (Pretrain + AT) (Singh et al., 2023) and Swin-B (MIMIR) (Xu et al., 2023). DDAttack consistently achieves higher ASR than the competing attacks throughout the query process on both models. Notably, its advantage appears early and persists as the query budget increases, indicating that distilled guidance enables more effective perturbations to be identified with fewer victim-model queries. These results further demonstrate the query efficiency of DDAttack beyond its final performance under the 5,000-query budget.

Average query count. Furthermore, Table 5 reports the average number of victim queries required for successful attacks. DDAttack consistently requires fewer queries than Square Attack across all evaluated victim models, showing that distilled guidance not only improves overall attack effectiveness but also enables successful adversarial examples to be found earlier in the search process.

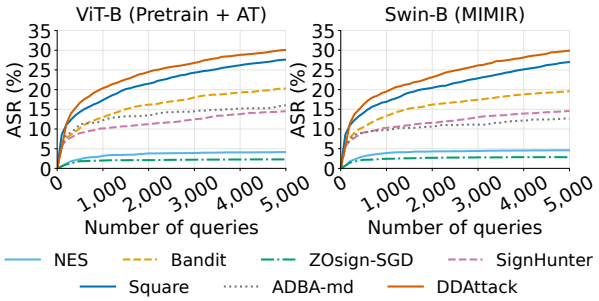


Figure 1: Attack success rate (ASR) versus the number of victim queries on ViT-B (Pretrain + AT) and Swin-B (MIMIR) under a 5,000-query budget and an ℓ_∞ perturbation constraint with $\epsilon = 8/255$.

Table 5: Average number of victim queries for successful attacks on ImageNet-1K under an ℓ_∞ perturbation constraint with $\epsilon = 8/255$, comparing Square Attack and DDAttack.

Victim model	Average query count	
	Square	DDAttack
ViT-S (Pretrain + AT)	1063.85	968.20
ViT-B (Pretrain + AT)	1091.46	1010.29
Swin-B (Input Grad Reg.)	1117.12	1019.86
Swin-L (Input Grad Reg.)	1114.70	1067.08
Swin-B (ARES-Bench AT)	1080.69	954.45
Swin-L (ARES-Bench AT)	1068.32	971.30
Swin-B (MIMIR)	1168.29	1102.80
Swin-L (MIMIR)	1152.73	1036.77

Table 6: Directional alignment with independently generated PGD perturbations under ℓ_∞ with $\epsilon = 8/255$ and a 5,000-query black-box budget. The paired gap denotes the per-input cosine-similarity difference between DDAttack and Square Attack.

Victim model	Square cosine	DDAttack cosine	Paired gap	95% CI
ViT-B (Pretrain + AT)	0.2452	0.2534	+0.0082	[0.0060, 0.0104]
Swin-B (MIMIR)	0.2648	0.2725	+0.0077	[0.0053, 0.0101]

Alignment with white-box perturbations. We further examine the efficiency of adversarial perturbations generated by DDAttack in terms of their alignment with white-box attack directions obtained by PGD attack (Madry et al., 2017). For each clean-correct input x_i, we independently generate a PGD perturbation δ_i^{PGD} under the same ℓ_∞ constraint ($\epsilon = 8/255$) and measure its cosine similarity c_i^A with the final perturbation produced δ_i^A by a black-box attack A:

$$c_i^A = \frac{\langle \delta_i^A, \delta_i^{\mathrm{PGD}} \rangle}{\|\delta_i^A\|_2 \|\delta_i^{\mathrm{PGD}}\|_2}. \tag{19}$$

The same PGD reference is used for both DDAttack and Square Attack on each input to ensure a paired comparison, where Square Attack corresponds to random initialization followed by random search. As shown in Table 6, under a budget of 5,000 queries, DDAttack achieves mean cosine similarities of 0.2534 and 0.2725 on ViT-B (Pretrain + AT) and Swin-B (MIMIR), respectively, compared with 0.2452 and 0.2648 for Square Attack. We further compute 95% confidence intervals for the paired differences in cosine similarity between DDAttack and Square Attack, as shown in the table. Since both intervals lie above zero, the results indicate a statistically significant positive increase in directional alignment. Additional comparisons with surrogate-based attacks, hyperparameter sensitivity analyses, and computational-cost evaluations are provided in the Appendix B.

5 Conclusion

In this work, we investigate dataset distillation as a new source of prior information for score-based black-box adversarial attacks and propose **DDAttack**. By exploiting the rich information encoded in distilled examples, DDAttack improves adversarial search through two complementary components: distillation-guided initialization, which provides a stronger starting point, and distillation-guided search, which uses distilled priors to guide subsequent perturbation updates. Importantly, DDAttack requires no auxiliary surrogate model during attack generation. Extensive experiments show that DDAttack consistently achieves stronger attacks than representative score-based baselines under fixed query budgets and different perturbation constraints. More broadly, our results demonstrate that distilled data can serve as an effective source of prior information for black-box adversarial attacks, extending their utility beyond conventional model training. We hope this finding sheds light on the broader potential of dataset distillation for adversarial learning and motivates future research on exploiting distilled knowledge for more efficient and effective black-box attacks.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code development and to improve the readability and language of the manuscript. We did not use generative AI tools to generate synthetic datasets; develop theoretical models or conceptual frameworks; formulate mathematical claims or proofs; propose or refine research hypotheses; design the research methodology or experiments; clean or reformat datasets; perform qualitative or thematic data analysis; or interpret experimental results. All AI-assisted manuscript edits were reviewed and revised by the authors, and all AI-assisted code was manually inspected, verified, and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies score-based adversarial attacks on image classification to better understand model vulnerabilities and support robustness evaluation. Our experiments use only ImageNet and do not involve additional human subjects or the collection of new personal data. We acknowledge that adversarial attack methods have potential for misuse, and we intend this work to support controlled security evaluation and the development of more robust models.

REPRODUCIBILITY STATEMENT

We include the source code of DDAttack in the supplementary material. All victim models and distilled datasets used in our experiments are publicly available through their official repositories, with details provided in the Appendix. For all baseline methods, we use the official implementations and report the corresponding implementation settings in the Appendix. The source code of DDAttack will be publicly released upon acceptance.

REFERENCES

Abdullah Al-Dujaili and Una-May O'Reilly. Sign bits are all you need for black-box attacks. In *International Conference on Learning Representations*, 2020.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pp. 484–501. Springer, 2020.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Generalizing dataset distillation via deep generative prior. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3739–3748, 2023.

Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *Advances in neural information processing systems*, 32, 2019.

Shuyu Cheng, Guoqiang Wu, and Jun Zhu. On the convergence of prior-guided zeroth-order optimization algorithms. *Advances in Neural Information Processing Systems*, 34:14620–14631, 2021.

Shuyu Cheng, Yibo Miao, Yinpeng Dong, Xiao Yang, Xiao-Shan Gao, and Jun Zhu. Efficient black-box adversarial attacks via bayesian optimization guided by a function prior. *arXiv preprint arXiv:2405.19098*, 2024.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

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 Datasets and Benchmarks Track*, 2021.

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

Jiawei Du, Yidi Jiang, Vincent YF Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3749–3758, 2023.

Yan Feng, Baoyuan Wu, Yanbo Fan, Li Liu, Zhifeng Li, and Shu-Tao Xia. Boosting black-box attack with partially transferred conditional adversarial distribution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15074–15083, 2022.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Yiwen Guo, Ziang Yan, and Changshui Zhang. Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. In *International Conference on Learning Representations*, 2020.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, pp. 2137–2146, 2018.

Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *International Conference on Learning Representations*, 2019.

Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *International Conference on Machine Learning*, pp. 11102–11118, 2022.

Shiye Lei and Dacheng Tao. A comprehensive survey of dataset distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):17–32, 2023.

Chang Liu, Yinpeng Dong, Wenzhao Xiang, Xiao Yang, Hang Su, Jun Zhu, Yuefeng Chen, Yuan He, Hui Xue, and Shibao Zheng. A comprehensive study on robustness of image classification models: Benchmarking and rethinking. *International Journal of Computer Vision*, 133(2):567–589, 2025.

Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. Signsgd via zeroth-order oracle. In *InternationalConferencee on Learning Representations*, 2019.

Nicholas A Lord, Romain Mueller, and Luca Bertinetto. Attacking deep networks with surrogate-based adversarial black-box methods is easy. *arXiv preprint arXiv:2203.08725*, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Adrián Rodríguez-Muñoz, Tongzhou Wang, and Antonio Torralba. Characterizing model robustness via natural input gradients. In *European Conference on Computer Vision*, pp. 161–178, 2024.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in neural information processing systems*, 33:3533–3545, 2020.

Naman Deep Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. *Advances in Neural Information Processing Systems*, 36:13931–13955, 2023.

Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9390–9399, 2024.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Yusuke Tashiro, Yang Song, and Stefano Ermon. Diversity can be transferred: Output diversification for white-and black-box attacks. *Advances in neural information processing systems*, 33:4536–4548, 2020.

Feiyang Wang, Xingquan Zuo, Hai Huang, and Gang Chen. Adba: Approximation decision boundary approach for black-box adversarial attacks. In *Proceedings of the AAAI conference on artificial intelligence*, 2025.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Xiaoyun Xu, Shujian Yu, Zhuoran Liu, and Stjepan Picek. Mimir: Masked image modeling for mutual information-based adversarial robustness. *arXiv preprint arXiv:2312.04960*, 2023.

Jiancheng Yang, Yangzhou Jiang, Xiaoyang Huang, Bingbing Ni, and Chenglong Zhao. Learning black-box attackers with transferable priors and query feedback. *Advances in Neural Information Processing Systems*, 2020.

Fei Yin, Yong Zhang, Baoyuan Wu, Yan Feng, Jingyi Zhang, Yanbo Fan, and Yujiu Yang. Generalizable black-box adversarial attack with meta learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1804–1818, 2023a.

Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):150–170, 2024. doi: 10.1109/TPAMI.2023.3323376.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*, 2020.

Meixi Zheng, Xuanchen Yan, Zihao Zhu, Hongrui Chen, and Baoyuan Wu. Blackboxbench: A comprehensive benchmark of black-box adversarial attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7867–7885, 2025.

A THEORETICAL MOTIVATION

We provide a local analysis to motivate why distilled references can improve query efficiency and how the main components of DDAttack contribute to the search. This analysis is organized around three questions: (i) how victim-independent distilled prior knowledge is constructed into victim-specific guidance, and (ii) why distillation-guided direction can make more consistent progress than unguided random search. The analysis rests on a first-order approximation of the victim margin and is intended to explain the mechanism of DDAttack rather than to establish a global convergence guarantee.

Local margin model. Recall that the untargeted margin of a perturbation δ is

$$m(\delta) = f_y(\widetilde{x}(\delta)) - \max_{k \neq y} f_k(\widetilde{x}(\delta)). \tag{20}$$

An attack succeeds when $m(\delta) < 0$. Around the current perturbation δ_t, and assuming the active competing class remains unchanged locally, we approximate

$$m(\delta_t + \zeta) \approx m(\delta_t) + \langle g_t, \zeta \rangle, \tag{21}$$

where $g_t = \nabla_\delta m(\delta_t) \in \mathbb{R}^{C\times H\times W}$ denotes the gradient of the victim margin with respect to the adversarial perturbation δ_t. Although g_t is not accessible by the attacker, Eq. 21 shows that a proposal ζ reduces the margin to first order exactly when $\langle g_t, \zeta\rangle < 0$. Query-efficient search can therefore be viewed as identifying proposal directions that align with the negative margin gradient, $-g_t$, without directly accessing g_t.

Distilled priors as directional guidance. The distilled references s_i are constructed independently of the victim model, so on their own they carry no guarantee of aligning with g_t. DDAttack instead evaluates the perturbations δ_i built from these references using victim feedback (Eq. 8) before using them for search guidance, and this screening step has a direct interpretation under the local model. Around the clean input, Eq. 21 gives

$$m_i = m(\delta_i) \approx m(0) + \langle g_0, \delta_i\rangle. \tag{22}$$

Since $m(0)$ is shared by all candidates, to first order, ranking candidates by m_i is equivalent to ranking them by $\langle g_0, \delta_i\rangle$: selecting the smallest-margin candidates approximately selects the references whose induced perturbations are best aligned with local margin reduction. The candidate evaluation in Sec. 3.2 can thus be viewed as a zeroth-order procedure for converting a victim-independent distilled prior into victim-specific directional guidance, using only $|\mathcal{S}(x)|$ queries.

The margin-based aggregation in Eq. 11 further emphasizes candidates with smaller victim margins. Specifically, the weights

$$w_i = \frac{e^{-m_i/\tau_d}}{\sum_{j\in\mathcal{I}_G} e^{-m_j/\tau_d}}, \tag{23}$$

are exactly the solution to the following entropy-regularized optimization problem:

$$\min_{\mathbf{w}\in\mathcal{W}_G} \sum_{i\in\mathcal{I}_G} w_i m_i + \tau_d \sum_{i\in\mathcal{I}_G} w_i \log w_i, \qquad \mathcal{W}_G = \left\{ \mathbf{w} : w_i \geq 0,\ \sum_{i\in\mathcal{I}_G} w_i = 1 \right\}. \tag{24}$$

Eq. 23 assigns greater weight to low-margin candidates, which provide stronger local evidence of alignment with margin reduction, while the entropy term discourages the solution from collapsing onto a single candidate. This reduces the risk of allowing the guidance direction to depend entirely on one potentially noisy estimate of the local descent direction. Combining Eq. 22 with Eq. 24, the resulting guidance direction $\bar{\delta} = \sum_i w_i \delta_i$ can therefore be interpreted as a margin-weighted aggregation of candidates that exhibit stronger local evidence of decreasing the victim margin.

Guided search versus random search. We next compare the proposal mechanisms of guided and random search. For clarity, we ignore the clipping operations that enforce the valid image range and the ℓ_∞ perturbation constraint in Eqs. 2 and 17. Under this simplification, a guided proposal can be written locally as

$$\zeta_t^{\text{guide}} = \eta\, P_t \odot q, \tag{25}$$

where η denotes the proposal magnitude 2ϵ as in Eq. 17, P_t is the patch mask, and $q = \operatorname{sign}(\bar{\delta})$ is the distilled guidance direction. We then define its patch-wise adversarial alignment as

$$A_t(P_t) = -\left\langle g_t, P_t \odot q \right\rangle, \tag{26}$$

so that $A_t(P_t) > 0$ indicates the guided proposal decreases the margin to first order. Because DDAttack accepts a proposal only when the margin decreases (Eq. 18), its realized first-order change is

$$\Delta_t^{\text{guide}} \approx -\eta\, [A_t(P_t)]_+, \qquad [z]_+ = \max(z, 0). \tag{27}$$

For the random search, DDAttack samples an independent sign $u_{t,c} \in \{-1, +1\}$ for each of the C channels, consistent with the per-channel sampling of u_t in Eq. 16. Define the gradient mass within the selected patch $\mathcal{R}_t$ for channel c as

$$h_{t,c}(P_t) = \sum_{(u,v)\in\mathcal{R}_t} g_{t,c,u,v}, \tag{28}$$

where $g_{t,c,u,v}$ denotes the (c, u, v) entry of g_t. The corresponding first-order margin change is

$$\Delta_t^{\text{rand}} \approx \eta \sum_{c=1}^{C} u_{t,c}\, h_{t,c}(P_t), \tag{29}$$

and by symmetry of the random signs, $\mathbb{E}_{u_t}[\Delta_t^{\mathrm{rand}} \mid P_t] = 0$ (i.e., a random proposal carries no systematic bias toward reducing the margin). Random search nonetheless makes progress because DDAttack rejects margin-increasing proposals, so only the favorable tail is realized. By symmetry,

$$\mathbb{E}_{u_t}\left[\min\{\Delta_t^{\mathrm{rand}}, 0\} \mid P_t\right] = -\frac{\eta}{2}\mathbb{E}_{u_t}\left|\sum_{c=1}^{C} u_{t,c}\, h_{t,c}(P_t)\right|. \tag{30}$$

Random search and guided search thus make progress through different mechanisms: random search extracts progress from the *variance* of Δ_t^{rand} via accept/reject filtering, while guided search extracts progress from the *systematic bias* $A_t(P_t)$ induced by the distilled prior.

Combining the two branches according to the mixture rule (Eq. 16), the expected first-order margin change under DDAttack's search is

$$\begin{aligned}\mathbb{E}[m_{t+1} - m_t] \approx -\eta\Bigg[&\gamma_t\, \mathbb{E}_{P_t}\left[A_t(P_t)_+\right] \\ &+ \frac{1-\gamma_t}{2}\mathbb{E}_{P_t,u_t}\left|\sum_{c=1}^{C} u_{t,c}\, h_{t,c}(P_t)\right|\Bigg].\end{aligned} \tag{31}$$

Under the local first-order approximation, distilled guidance provides greater expected progress than pure random search when

$$\mathbb{E}_{P_t}\left[A_t(P_t)_+\right] > \frac{1}{2}\mathbb{E}_{P_t,u_t}\left|\sum_{c=1}^{C} u_{t,c}\, h_{t,c}(P_t)\right|. \tag{32}$$

This condition is equivalent to requiring the guided direction to provide, on average, a larger reduction in the victim margin than the favorable random proposals. Increasing the size of the distilled reference set $\mathcal{S}(x)$ increases the likelihood of satisfying this condition. Specifically, suppose that a perturbation constructed from a randomly sampled distilled reference has probability $p > 0$ of exhibiting sufficiently strong margin-reducing alignment. For $|\mathcal{S}(x)|$ independently sampled references, the probability that at least one candidate exhibits such alignment is

$$1 - (1-p)^{|\mathcal{S}(x)|}, \tag{33}$$

which increases monotonically with N. Since DDAttack evaluates these candidates using the victim margin and assigns larger weights to lower-margin candidates, a sufficiently informative and larger candidate set is more likely to produce a guidance direction that satisfies Eq. 32. This assumption is also supported by the empirical trend observed in Fig. 2.

Discussion In this theoretical analysis, we show that DDAttack can produce more efficient proposals than random selection, which is further supported by the PGD-direction alignment experiment in Sec. 4.4. The distilled guidance direction shows stronger alignment with the white-box margin-descent direction than random proposals, consistent with Eq. 32.

B Extra experiments

Comparison with surrogate-based attacks. We further compare DDAttack with P-RGF (Cheng et al., 2019) and SimBA-ODS (Tashiro et al., 2020), two representative surrogate-based black-box attacks, under the same 5,000-query budget and ℓ_∞ perturbation budget of $\epsilon = 8/255$. For both baselines, we retain their released ImageNet hyperparameter settings and evaluate them using either a non-robust ResNet-151 or a robust ResNet-50 (Salman et al., 2020) as the surrogate model. The results show that surrogate-based performance is sensitive to the choice of surrogate, since the effectiveness of the transferred guidance depends on how well the surrogate aligns with the victim model. In contrast, DDAttack requires no surrogate model and achieves the lowest robust accuracy across all eight victim models. As shown in Table 7, compared with the strongest surrogate-based result for each victim model, DDAttack further reduces robust accuracy by 5.40 to 9.54 percent, with an average reduction of 7.26 percent. These results demonstrate that distilled examples can provide effective prior information for black-box adversarial search without relying on the transferability of an auxiliary surrogate model.

Table 7: Comparison with surrogate-based black-box attacks on ImageNet-1K under ℓ_∞ perturbations with $\epsilon = 8/255$ and a 5,000-query budget. For P-RGF and SimBA-ODS, we report results using either a non-robust ResNet-151 or a robust ResNet-50 (Salman et al., 2020) as the surrogate model, while DDAttack requires no surrogate model. Lower robust accuracy indicates a stronger attack.

Victim model	Clean Acc.	P-RGF		SimBA-ODS		DDAttack
		R151 (non-robust)	R50 (robust)	R151 (non-robust)	R50 (robust)	
ViT-S (Pretrain + AT)	72.58	70.82	57.50	68.70	53.96	47.34
ViT-B (Pretrain + AT)	76.30	74.26	64.58	72.42	60.34	53.36
Swin-B (Input Grad Reg.)	77.76	72.44	66.94	71.98	61.86	54.96
Swin-L (Input Grad Reg.)	79.36	75.38	69.94	74.22	63.04	57.64
Swin-B (ARES-Bench AT)	76.16	71.72	66.06	69.30	62.54	54.98
Swin-L (ARES-Bench AT)	78.92	74.60	69.32	73.94	66.82	58.64
Swin-B (MIMIR)	76.62	73.38	67.08	72.76	63.26	53.72
Swin-L (MIMIR)	78.62	74.96	69.92	74.16	65.14	58.28

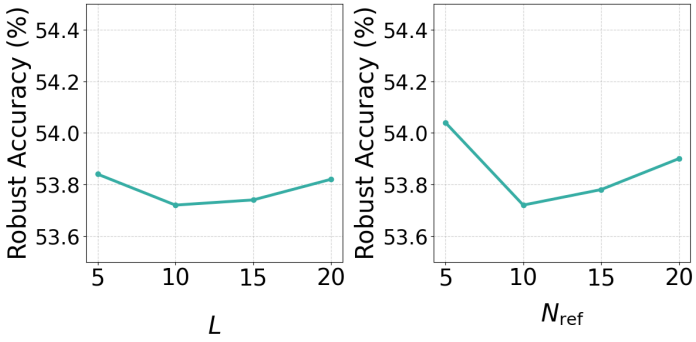


Figure 2: Sensitivity to the number of competing classes L and distilled references per class N_{ref} on Swin-B (MIMIR) under ℓ_∞ perturbations with $\epsilon = 8/255$. The total query budget is fixed at 5,000. Increasing L or N_{ref} enlarges the reference set $\mathcal{S}(x)$.

Effect of reference-set size. We further study the sensitivity of DDAttack to the number of competing classes L and the number of distilled references per class N_{ref}. As shown in Fig. 2, when L or N_{ref} is too small, the reference set may provide insufficient distilled guidance for constructing effective candidate perturbations. In contrast, increasing L or N_{ref} enlarges $\mathcal{S}(x)$ and therefore consumes more of the query budget during initialization, leaving fewer queries for the subsequent search stage. The best performance is obtained around $L = 10$ and $N_{\text{ref}} = 10$, while further increasing the reference-set size does not improve attack effectiveness. These results suggest that a moderate number of distilled references $\mathcal{S}(x)$ provides a favorable balance between initialization and sufficient budget for iterative adversarial search.

Hyperparameter sensitivity test. We further examine the sensitivity of DDAttack to the hyperparameters controlling the adaptive search scale, including the margin threshold θ, temperature τ_p, and minimum patch-area fraction $p_{\min}$. We vary one hyperparameter at a time while keeping all others fixed, and evaluate robust accuracy on Swin-B (MIMIR) under an ℓ_∞ perturbation budget of $\epsilon = 8/255$.

As shown in Fig. 3, DDAttack remains relatively stable across a broad range of hyperparameter values. For θ, robust accuracy changes only slightly when θ varies from 1 to 5, while larger values lead to a noticeable degradation. For τ_p, the best performance is obtained around $\tau_p = 0.5$, whereas overly small or large values reduce attack effectiveness. Similarly, $p_{\min} = 0.075$ achieves the lowest robust accuracy, with only modest variations for nearby values. Overall, the results indicate that DDAttack is not highly sensitive to these hyperparameters within reasonable ranges, while the default settings $\theta = 5$, $\tau_p = 0.5$, and $p_{\min} = 0.075$ are used for consistency. For $p_{\max}$, we follow the default value in Square attack and set it to 0.8.

Hardware and computation. Table 8 reports the average attack time per image on a single NVIDIA H100 GPU. Despite introducing distillation-guided initialization and search, DDAttack incurs only

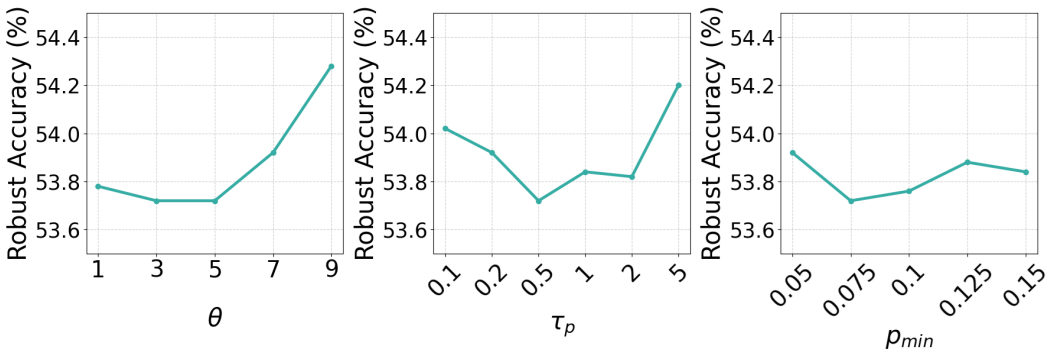


Figure 3: Hyperparameter sensitivity of DDAttack on Swin-B (MIMIR) under an ℓ_∞ perturbation budget of $\epsilon = 8/255$. We report the robust accuracy with different θ, τ_p, and $p_{\min}$ individually while keeping the remaining hyperparameters fixed.

Table 8: Average attack runtime (seconds per image) on a single NVIDIA H100 GPU under a 5,000-query budget. For all attack methods, we disable early termination and run exactly 5,000 queries per image. Results are averaged over all evaluated images. Lower is better.

Victim model	NES	Bandit	ZOsign	SignHunter	Square	ADBA-md	DDAttack
ViT-S (Pretrain + AT)	2.46	2.41	2.82	2.34	2.31	18.81	2.33
ViT-B (Pretrain + AT)	5.78	5.19	6.16	5.22	4.90	18.51	5.41
Swin-B (Input Grad Reg.)	6.72	6.02	7.65	5.97	6.54	56.57	6.47
Swin-L (Input Grad Reg.)	12.76	10.93	13.18	11.27	10.35	73.07	10.53
Swin-B (ARES-Bench AT)	6.83	5.92	7.32	6.00	6.64	52.38	6.53
Swin-L (ARES-Bench AT)	12.47	10.81	13.27	11.26	10.66	70.94	10.53
Swin-B (MIMIR)	6.51	5.99	7.10	6.03	6.39	59.50	6.63
Swin-L (MIMIR)	13.95	12.40	14.71	12.98	11.83	61.12	11.97

limited additional computational overhead compared with Square Attack. Averaged across the eight victim models, the runtime of DDAttack remains comparable to other lightweight score-based attacks and is substantially lower than ADBA-md.

C Extra implementation details

Baseline implementations and hyperparameters. We use the implementations of NES (Ilyas et al., 2018), Bandit (Ilyas et al., 2019), ZOsign-SGD (Liu et al., 2019), and SignHunter (Al-Dujaili & O'Reilly, 2020) provided by BlackboxBench[1] (Zheng et al., 2025), and obtain ADBA-md (Wang et al., 2025) from the authors' official repository[2]. Our Square baseline uses the implementation adapted from Square Attack (Andriushchenko et al., 2020) implemented in AutoAttack[3] (Croce & Hein, 2020). We adopt published ImageNet hyperparameters where available and otherwise use released default configurations with the required input-scale conversions. All methods share victim preprocessing and the ImageNet evaluation subset, as standardized in robustbench (Croce et al., 2021).

Distilled dataset preparation. We use distilled datasets produced by RDED (Sun et al., 2024) and SRe2L (Yin et al., 2023b) in our experiments. For both methods, we directly adopt the officially released distilled samples provided by the respective official implementation, without regenerating or further optimizing them. These distilled datasets are generated from the ImageNet-1K training set and are used as fixed prior information throughout attack generation.

[1] https://github.com/SCLBD/BlackboxBench.
[2] https://github.com/BUPTAIOC/ADBA.
[3] https://github.com/fra31/auto-attack.

DDAttack patch-size schedule. As mentioned in Sec. 3, for an image in patch search, we set its initial patch-area parameter once, using

$$p_0 = p_{\min} + (p_{\max} - p_{\min}) \operatorname{sigmoid}\left(\frac{m_0 - \theta}{\tau_p}\right), \tag{34}$$

where $p_{\min} = 0.075$, $p_{\max} = 0.8$, $\theta = 5$ and $\tau_p = 0.5$. Thus, smaller initial margins produce finer initial patches, while larger margins produce coarser patches. The initial value is held fixed as the reference scale throughout the subsequent search.

Let $t = \{0, 1, \ldots, T-1\}$ index patch proposals after DD initialization, where $T = Q - |\mathcal{S}(x)| - 1$ is the allocated patch-search budget. We use the piecewise halving schedule

$$\begin{aligned} p_t &= p_0 \, 2^{-h(t)}, \\ h(t) &= \sum_{b \in \mathcal{B}} \mathbf{1}\{t > b\}, \end{aligned} \tag{35}$$

with milestone

$$\mathcal{B} = \{10, 50, 200, 500, 1000, 2000, 4000\}.$$