

On the Quantum Advantage in Black-Box Adversarial Attacks (Extended Abstract)

Fei Meng, Junhao Xu, Wanli Cheng, Yunlong Xiao, Jinge Bao, Mert D. Pesé, Siyu Huang

1 Introduction

Deep neural networks remain vulnerable to adversarial examples—small input perturbations that flip a prediction. In the practical *black-box* threat model the attacker can only query the model’s outputs, with no gradient access [ACFH20, CJW20, VAR24]. Early single-pixel [SVS19] and multi-pixel [CH19, MMDF19] attacks exposed classifier vulnerability to sparse perturbations, and later methods improved query efficiency via random [CAS⁺22] or Bayesian [VAR24] search. This is especially hard for sparse ℓ_0 attacks that modify only a few pixels, requiring search through a vast discrete space; state-of-the-art methods still need 10^2 – 10^4 queries per image [CAS⁺22, VAR24], which scales poorly with resolution and makes generating enough samples for robust training prohibitively expensive. We adopt Sparse-RS [CAS⁺22] as our baseline because its query mechanism—one evaluation per query, no amortized state—aligns exactly with our oracle model; other families optimizing different norms or feedback types (score-based [ACFH20], label-only [CJW20]) are not directly comparable on ℓ_0 query efficiency [ZZY⁺25].

Quantum search subroutines provide proven speedups for unstructured search [Gro96, BBBV97], yet existing quantum adversarial ML work overwhelmingly studies the *reverse* direction—the vulnerability of quantum classifiers to classical attacks [LDD20, LW20, WEL⁺23, TYZ⁺25], robustness certificates [DWS⁺26], and quantum-enhanced defenses [WTL⁺23]; to our knowledge no prior work attacks classical models with quantum search. By encoding the pixel-selection and color-substitution process into a quantum register and treating the black-box classifier as a quantum oracle—so that one quantum query is methodologically equivalent to one classical evaluation—Grover’s algorithm [Gro96] promises a quadratic query speedup. However, its optimal iteration count depends on the number of successful perturbations M , which is unknown *a priori* and varies across inputs; estimating M via quantum counting introduces a substantial $\Omega(\sqrt{N})$ up-front cost, creating a regime in which the overhead alone exhausts the attacker’s budget. Variants such as BBHT [BBHT98] search without knowing M , and threshold-tightening yields quantum minimum-finding (GAS) [DH96, Bul07], but their application to ML security remains largely unexplored.

Our contributions are: (i) the first formalization of black-box adversarial attacks as a quantum oracle search problem, with a proof of quantum advantage over classical methods; (ii) a Grover Adaptive Search (GAS) framework that eliminates the counting bootstrap and delivers a practical speedup under realistic query constraints; and (iii) a closed-form characterization of the quantum-advantage crossover, validated on datasets from small digits to ImageNet.

2 Main Results

Problem setting. We cast the sparse ℓ_0 attack as discrete search. A candidate k -pixel perturbation is a bitstring $z \in \{0, 1\}^n$; for single-pixel attacks ($k=1$), z concatenates a pixel position p (using a Morton Z-order encoding to preserve spatial locality under the Hamming metric) and a binarized color c (Gray-coding the RGB-cube corners). The perturbed image is $x(z) = x_0 \oplus z$, and the search space has size $N := 2^n$. Given a clean image x_0 correctly predicted as $y_0 = f(x_0)$, we define the adversarial energy via the logit margin

$$E(z) := \ell_{y_0}(x_0 \oplus z) - \max_{c \neq y_0} \ell_c(x_0 \oplus z), \quad (1)$$

where ℓ denotes pre-softmax scores. An (untargeted) attack succeeds iff $E(z) \leq 0$. The number of solutions is $M := |\{z : E(z) \leq 0\}|$ and their density $\rho := M/N$. In our black-box oracle model, querying $E(z) \leq 0$

for one candidate costs exactly one oracle query, matching a single classical forward pass; attack efficiency is measured by the *first-hit budget* T .

We compare against Sparse-RS [CAS⁺22], a strong score-based classical baseline whose single-forward-pass query mechanism aligns exactly with our oracle model, enabling an oracle-honest head-to-head comparison (one query = one call to $E(z)$).

Lemma 1 (Sparse-RS single-pixel first-hit). *For single-pixel attacks, if $N > M \gg 2^{n_{\text{color}}}$ and density $\rho = M/N$, then $\mathbb{E}[T_{\text{first-hit}}^{\text{SR}}] = (1/\rho)(1 + O(2^{n_{\text{color}}}/M))$.*

Classical search thus scales linearly as $1/\rho$: to leading order the baseline needs $1/\rho$ queries on average.

Theorem 1 (Grover dead zone). *In the single-pixel asymptotic regime, for error tolerance ϵ , the counting phase of vanilla Grover costs $T_{\text{count}}(\rho, \epsilon) := \pi/(\epsilon\sqrt{\rho})$. For any budget $T < T_{\text{count}}(\rho, \epsilon)$, vanilla Grover deterministically finds zero successful attacks, whereas Sparse-RS finds at least one with probability $1 - (1 - \rho)^T + o(1)$.*

Theorem 2 (Quantum advantage over Sparse-RS). *In the same regime: (1) under BBHT, the GAS first-hit budget satisfies $\mathbb{E}[T_{\text{first-hit}}^{\text{GAS}}] \leq 4.5/\sqrt{\rho}$ with no counting bootstrap, and asymptotically outperforms Sparse-RS for densities below the leading-order threshold $\rho < 4/81 \approx 4.9\%$; (2) vanilla Grover under amplitude estimation outperforms Sparse-RS only below a much sparser threshold, delayed by its $\sqrt{N}$ bootstrap penalty.*

Because classical search scales linearly with $1/\rho$ while quantum search scales with $1/\sqrt{\rho}$, quantum methods inevitably win as attacks become harder (i.e., sparser). GAS-BBHT enjoys this scaling immediately by bypassing the counting phase, whereas vanilla Grover realizes it only on much harder targets where its initial overhead can be amortized.

3 Main Techniques

Grover search initializes a uniform superposition over all N candidates and iterates the Grover operator $G = DO$: a phase oracle $O|z\rangle = (-1)^{\mathbf{1}_{\{E(z) \leq 0\}}} |z\rangle$ flips the sign of successful states (one application = one classical-equivalent query), and a diffusion operator D redistributes probability mass onto them. After $k^* \approx (\pi/4)\sqrt{N/M}$ iterations, the success probability is maximized. Because M is unknown and varies across images, applying G too few or too many times “under-” or “over-shoots” the amplitude. Vanilla Grover must therefore prepend a quantum counting subroutine [BHMT00] to estimate M ; this bootstrap costs $\Omega(\sqrt{N})$ and, under a tight budget, consumes all queries before any search begins.

To bypass the bootstrap, we build on Grover Adaptive Search [DH96, BBW05, Bul07] driven by the BBHT schedule [BBHT98]. Rather than pre-computing M , BBHT draws an iteration count k_t uniformly from a dynamically growing window $[0, m_t - 1]$; on a failed measurement the window expands ($m_{t+1} = \min(\lceil 1.2 m_t \rceil, \lceil \sqrt{N} \rceil)$) and retries. This randomized strategy achieves the optimal $4.5\sqrt{N/M}$ query complexity while requiring only $O(1)$ first-round cost, so the quantum advantage materializes even under strict budgets. We wrap BBHT in a *threshold-tightening* loop that globally samples the improving subset $\{z : E(z) < \tau\}$, updating $\tau \leftarrow E(z')$ on each better candidate, with a domain-specific *floor-cap* at the success boundary ($\tau=0$): since any $E(z) \leq 0$ already flips the prediction, we stop tightening there and redirect the speedup toward discovering *diverse* adversarial examples. Query accounting stays fully transparent: the total cost is just the cumulative Grover iterations plus measurement verifications, with no hidden calibration overhead.

4 Experiments

We benchmark across 11 image datasets spanning four spatial scales and both standard and adversarially-trained classifiers: UCI [AK98] (8×8) and USPS [Hul02] (16×16) digits against MLPs; CIFAR-10 [KH⁺09] (32×32) against ResNet-20 and an ℓ_∞ -robust WRN-28-10 [WPD⁺23]; Tiny-ImageNet [LY⁺15] against a ResNet-50; and ImageNet [RDS⁺15] (256×256) against a standard ResNet-50, a robust ResNet-50 [SIE⁺20], and a ViT-B/16 [DBK⁺20] ($n=19$ qubits). Positions use a Morton space-filling curve [Mor66] and colors are endpoint-quantized, matching the classical baseline. All three attackers—Sparse-RS, vanilla Grover, and GAS-BBHT (ours)—run under *oracle-honest accounting*: every $E(z)$ call, Grover iteration, and post-measurement verification counts as one query. For each landscape, we evaluate the first 50 correctly

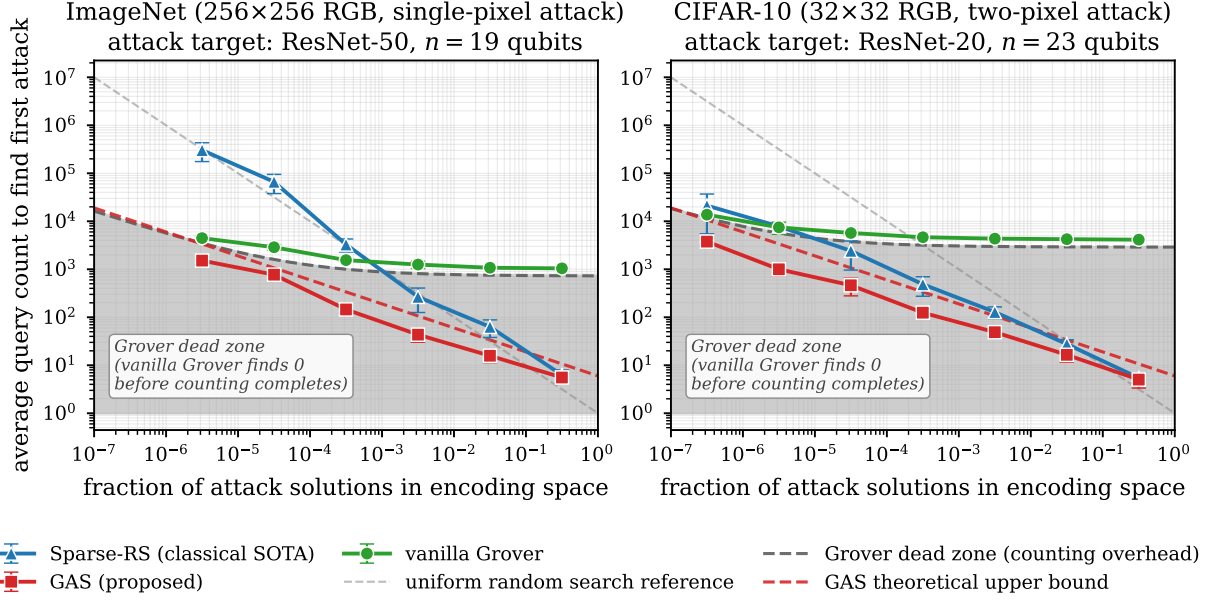


Figure 1: First-hit oracle cost versus marked density ρ for ImageNet single-pixel (left, $n=19$) and CIFAR-10 two-pixel (right, $n=23$) attacks. Solid: mean $\pm 95\%$ CI for Sparse-RS (blue $\triangle$), GAS-BBHT (red $\square$), vanilla Grover (green $\circ$). Dashed: $1/\rho$ uniform reference (grey), the Grover counting threshold $\pi/(\epsilon\sqrt{\rho})$ (green), and the GAS upper bound $4.5/\sqrt{\rho}$ (red). Shaded: the dead zone, inside which vanilla Grover finds nothing.

classified images with $\rho > 0$, computing the exact density $\rho := M/2^n$ by exhaustive enumeration (feasible since $n \leq 23$) and averaging over 10 independent chains per image; the resulting densities span seven decades ($\rho \in [\sim 10^{-7}, 0.5]$) and are aggregated into $\log_{10}$ -decade bins with 95% confidence intervals.

Empirical first-hit costs conform to the theory (Fig. 1). GAS-BBHT is the most efficient attacker in every density bin, operating 2–4 $\times$ below the $4.5/\sqrt{\rho}$ upper bound and tightly tracking the classical $1/\rho$ curve from below. As predicted by Theorem 2, GAS-BBHT strictly overtakes Sparse-RS once $\rho \lesssim 4/81$. Vanilla Grover also eventually overtakes the classical baseline, but its $\sqrt{N}$ bootstrap delays this crossover to much sparser densities ($\rho \approx 10^{-3}$). Across 400 per-image benchmarks, GAS-BBHT achieves a median speedup of 6.0 $\times$ on the first attack, *growing* to 9.6 $\times$ on a second distinct perturbation because it carries no estimation phase to amortize. Vanilla Grover’s counting cost is instead a one-time expense: it wins only 49% of first-attack cells (the dead zone), but its win rate jumps to 93% on the second attack (median 9.9 $\times$) once the marked-count estimate is cached.

On two-pixel CIFAR-10 ($n=23$) the classical baseline is no longer naive—Sparse-RS drops below $1/\rho$ by exploiting local spatial correlations—yet GAS-BBHT still keeps a decisive 2–4 $\times$ lead across the sparse regime ($\rho \in [10^{-5}, 10^{-2}]$). Extending the budget to $T=10^5$ exposes an amortization crossover: once vanilla Grover pays its one-time counting penalty, it overtakes GAS-BBHT on mid-density bins, whereas classical search suffices only in the ultra-dense regime ($\rho \geq 10^{-1}$).

Simulating symmetric depolarizing noise after each Grover iteration at per-qubit rates $p \in \{10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$ (matching near-term trapped-ion and superconducting hardware [Pre18, BCLK⁺22]), GAS-BBHT retains its quantum advantage ($r \geq 1$) across all benchmarks, with noisy query costs within a factor ~ 2 of ideal execution. Vanilla Grover’s dense-regime drop is driven by its counting overhead rather than noise amplification.

Feeding the discovered samples into a closed-loop training pipeline, at a constrained budget of $T=100$ queries per image, GAS-BBHT improves robust accuracy by +2.2 pp over Sparse-RS, driven by broader coverage of the target set. Vanilla Grover finds *zero* examples at $T=100$ (counting alone costs ~ 142 queries) and collapses to the weak baseline—a direct materialization of the dead zone. Clean accuracy stays within 0.6 pp of the original across all fine-tuned models.

References

- [ACFH20] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *Proceedings of the European Conference on Computer Vision*, pages 484–501, 2020.
- [AK98] E. Alpaydin and C. Kaynak. Optical Recognition of Handwritten Digits. UCI Machine Learning Repository, 1998. DOI: <https://doi.org/10.24432/C50P49>.
- [BBBV97] Charles H. Bennett, Ethan Bernstein, Gilles Brassard, and Umesh Vazirani. Strengths and weaknesses of quantum computing. *SIAM Journal on Computing*, 26(5):1510–1523, 1997.
- [BBHT98] Michel Boyer, Gilles Brassard, Peter Høyer, and Alain Tapp. Tight bounds on quantum searching. *Fortschritte der Physik: Progress of Physics*, 46(4-5):493–505, 1998.
- [BBW05] William P Baritompa, David W Bulger, and Graham R Wood. Grover’s quantum algorithm applied to global optimization. *SIAM Journal on Optimization*, 15(4):1170–1184, 2005.
- [BCLK⁺22] Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S Kottmann, Tim Menke, et al. Noisy intermediate-scale quantum algorithms. *Reviews of Modern Physics*, 94(1):015004, 2022.
- [BHMT00] Gilles Brassard, Peter Hoyer, Michele Mosca, and Alain Tapp. Quantum amplitude amplification and estimation. *arXiv preprint quant-ph/0005055*, 2000.
- [Bul07] David W. Bulger. Combining a local search and grover’s algorithm in black-box global optimization. *Journal of Optimization Theory and Applications*, 133(3):289–301, 2007.
- [CAS⁺22] Francesco Croce, Maksym Andriushchenko, Naman D Singh, Nicolas Flammarion, and Matthias Hein. Sparse-RS: A versatile framework for query-efficient sparse black-box adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6437–6445, 2022.
- [CH19] Francesco Croce and Matthias Hein. Sparse and imperceivable adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4724–4732, 2019.
- [CJW20] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *Proceedings of the 2020 IEEE Symposium on Security and Privacy*, pages 1277–1294, 2020.
- [DBK⁺20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [DH96] Christoph Durr and Peter Hoyer. A quantum algorithm for finding the minimum. *arXiv preprint quant-ph/9607014*, 1996.
- [DWS⁺26] Neil Dowling, Maxwell T. West, Angus Southwell, Azar C. Nakhel, Martin Sevier, Muhammad Usman, and Kavan Modi. Adversarial robustness guarantees for quantum classifiers. *npj Quantum Information*, 2026.
- [Gro96] Lov K Grover. A fast quantum mechanical algorithm for database search. In *Proceedings of the twenty-eighth Annual ACM Symposium on Theory of Computing*, pages 212–219, 1996.
- [Hul02] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 2002.

- [KH⁺09] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.
- [LDD20] Sirui Lu, Lu-Ming Duan, and Dong-Ling Deng. Quantum adversarial machine learning. *Physical Review Research*, 2(3):033212, 2020.
- [LW20] Nana Liu and Peter Wittek. Vulnerability of quantum classification to adversarial perturbations. *Physical Review A*, 101(6):062331, 2020.
- [LY⁺15] Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.
- [MMDF19] Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Sparsefool: a few pixels make a big difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9087–9096, 2019.
- [Mor66] Guy M. Morton. *A computer oriented geodetic data base and a new technique in file sequencing*. International Business Machines Company, 1966.
- [Pre18] John Preskill. Quantum computing in the nisc era and beyond. *Quantum*, 2:79, 2018.
- [RDS⁺15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [SIE⁺20] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pages 3533–3545, 2020.
- [SVS19] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.
- [TYZ⁺25] Dong Tan, Lili Yan, Jiayu Zhao, Yan Chang, and Shibin Zhang. A black-box attack method of machine learning algorithms based on quantum autoencoders. *Physica A: Statistical Mechanics and its Applications*, page 131033, 2025.
- [VAR24] Viet Quoc Vo, Ehsan Abbasnejad, and Damith C. Ranasinghe. Brusleattack: A query-efficient score-based black-box sparse adversarial attack. *arXiv preprint arXiv:2404.05311*, 2024.
- [WEL⁺23] Maxwell T. West, Sarah M. Erfani, Christopher Leckie, Martin Sevier, Lloyd C. L. Hollenberg, and Muhammad Usman. Benchmarking adversarially robust quantum machine learning at scale. *Physical Review Research*, 5(2):023186, 2023.
- [WPD⁺23] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *Proceedings of the International Conference on Machine Learning*, pages 36246–36263, 2023.
- [WTL⁺23] Maxwell T. West, Shu-Lok Tsang, Jia S. Low, Charles D. Hill, Christopher Leckie, Lloyd C. L. Hollenberg, Sarah M. Erfani, and Muhammad Usman. Towards quantum enhanced adversarial robustness in machine learning. *Nature Machine Intelligence*, 5(6):581–589, 2023.
- [ZYZ⁺25] 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*, 2025.