Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ27247

The Mechanisms, Evaluation, and Future Prospects of Black-Box Adversarial Attacks

Tianyi Huang

Shenzhen University, Shenzhen, China
2022271051@email.szu.edu.cn

Abstract. As machine learning is increasingly applied in healthcare, finance, and security, black-box adversarial attacks that manipulate inputs to mislead models without accessing internal details pose a growing security threat in real-world applications. This paper reviews black-box adversarial attacks by categorizing perturbation generation methods, such as gradient estimation and evolutionary search, along with attack targets, including directional and non-directional types. Besides, it distinguishes attack scenarios based on the attacker's knowledge into white-box and black-box categories, analyzing their effects across different machine learning tasks and summarizing related challenges and opportunities. The results indicate that in image classification tasks, zero-order optimization (ZOO) demands a large number of queries to reach a high success rate, whereas evolutionary algorithms enhance defense effectiveness by about 30% but involve greater computational costs. For large-scale language models, template injection attacks achieve success rates greater than 80%, while the PathSeeker multi-agent reinforcement learning framework achieves over 80% detection evasion by dynamically adjusting templates. However, the effectiveness of cross-domain transfer is still severely constrained. Key directions include meta-learning with Bayesian optimization, cross-modal attack frameworks, and testing standards for high-risk settings.

Keywords: Black-box adversarial attacks, Perturbation generation mechanisms, Adversarial machine learning, Security vulnerabilities

1. Introduction

In recent years, machine learning (ML) has advanced rapidly and is now used in critical areas like healthcare, finance, and security. However, these systems are still vulnerable to adversarial attacks that subtly alter inputs to fool predictions. Black-box attacks are especially concerning because they work without access to a model's internal details, making them realistic threats. Although methods like gradient estimation, surrogate models, and query-based optimization exist, much research still focuses on narrow tasks or ideal conditions, leaving gaps in understanding how black-box attacks work in real settings and limiting practical defenses. Based on existing literature, this paper explores the principles, technical approaches, and practical performance of black-box adversarial attacks, with a focus on how these attacks operate without access to internal model details. In particular, it analyzes their effectiveness in tasks such as image classification and natural language processing, and discusses the trade-offs between success rate, query cost, and detectability. Besides, it reviews

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.BJ27247

major attack types and methods, highlights their applicability across data modalities, and addresses challenges in balancing attack efficiency, query cost, as well as stealth. Accordingly, it highlights open issues requiring further research, aiming to guide efforts to enhance the security of AI systems deployed in sensitive environments [1].

2. Types and technical approaches of black-box attacks

2.1. The classification and technical principles of black-box adversarial attack methods

Black-box adversarial attacks encompass five main types, namely gradient estimation, evolutionary algorithms, decision-based methods, transfer attacks, and hybrid strategies. Each type has distinct characteristics in how perturbations are generated and how they interact with the target model, collectively forming the technical foundation of this field.

In particular, gradient estimation methods approximate the gradient direction by applying small perturbations to inputs and observing output changes [2]. This enables effective probing of model sensitivity and generalization across tasks. In contrast, evolutionary algorithms take inspiration from biological evolution and optimize adversarial samples through iterative procedures, including initialization, fitness evaluation, mutation, and selection [3]. As they do not depend on gradient or confidence scores, these algorithms are particularly effective in navigating complex search spaces. As shown in Figure 1, the general workflow includes evaluation, variation, and reinsertion steps to evolve fitter solutions over generations [4]. Decision-based methods, by contrast, modify inputs based solely on final output labels, making them applicable to black-box systems that return only hard labels, though often at the cost of high query consumption. To reduce query overhead, transfer attacks exploit the cross-model transferability of adversarial examples by crafting perturbations on surrogate models with similar architecture or data, then applying them to the target model [5]. In addition, hybrid strategies combine multiple methods to balance efficiency and robustness. For example, initial perturbations are generated via transfer attacks and then refined with evolutionary algorithms or gradient-free optimization [6].

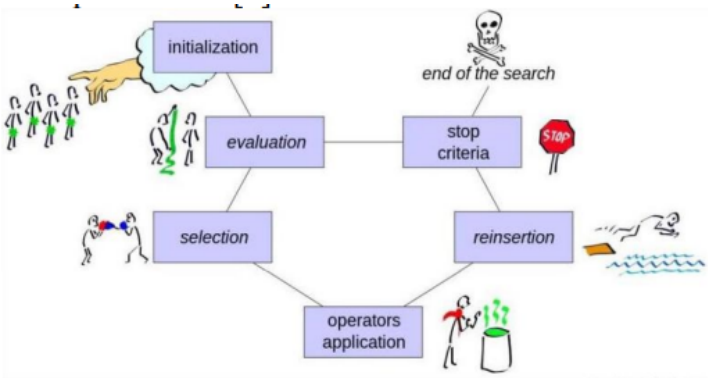


Figure 1. General schema of evolutionary algorithm

2.2. Attack scenarios based on adversarial knowledge transparency

Black-box attacks vary based on the attacker's knowledge transparency of the target model, directly influencing the feasibility, efficiency, and stealth of attack strategies. In zero-knowledge scenarios, attackers do not have access to the model's architecture, parameters, or training data and can only interact with the model through input-output queries, observing its predictions for specific inputs. In

DOI: 10.54254/2755-2721/2025.BJ27247

such cases, attacks often rely on trial and error, gradually adjusting inputs to cause misclassification, which is common in real-world scenarios. If attackers know the model type such as convolutional neural networks (CNNs) or training data categories, they can craft more targeted perturbations, thus improving efficiency while preserving the black-box assumption [7]. Furthermore, with greater knowledge of the target, attackers can build surrogate models to mimic its behavior. Perturbations crafted on the surrogate can then be transferred to the target, enabling transfer attacks that exploit shared vulnerabilities and perform well with partial knowledge. By relying solely on output labels, decision-based attackers iteratively modify inputs to cause misclassification, as shown in Figure 2 [8]. Besides, attackers respond to defensive measures by generating natural perturbations to bypass input filtering or randomization. In addition, the input modality like image or text affects both the required knowledge and attack strategy, increasing the complexity of black-box scenarios.

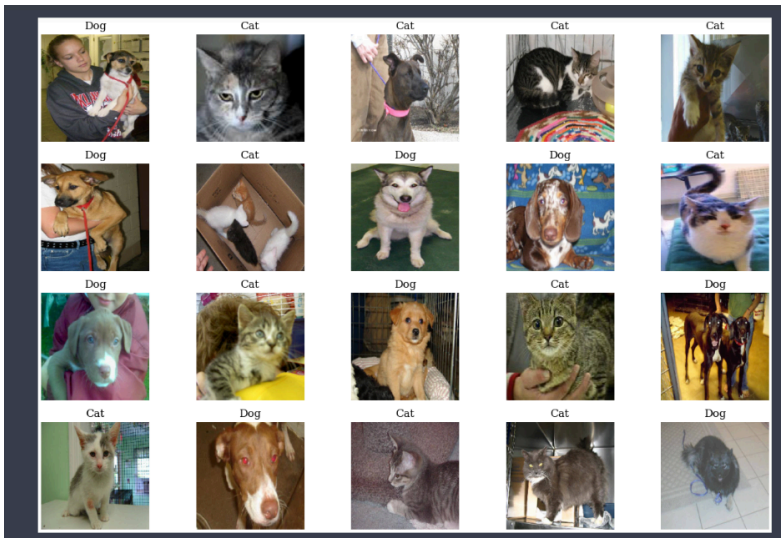


Figure 2. General schema of evolutionary algorithm

3. Evaluation and challenges of black-box attacks

3.1. Major evaluation metrics for black-box attacks

The effectiveness of black-box adversarial attacks is typically evaluated in terms of three major aspects, which are the attack success rate, query efficiency, and perturbation quality. In particular, the Attack Success Rate (ASR) measures whether the perturbation can effectively induce the model to produce incorrect predictions. Formally, it is defined as follows.

$$ASR = \frac{N_{success}}{N_{attak}} \tag{1}$$

This means that as ASR increases, the attack's effectiveness improves. Targeted attacks, which aim for specific incorrect outputs, generally present more challenges than untargeted ones, yet pose greater security risks. In typical image classification tasks, the ASR often ranges from 70% to 95%. A higher success rate indicates a stronger capability to mislead the target model [9].

The number of queries required to craft an adversarial example defines query efficiency. Since excessive queries hinder black-box attacks, maintaining a balance between attack success and query budget is essential. Formally, query efficiency is defined as follows.

DOI: 10.54254/2755-2721/2025.BJ27247

$$Avg_Queries = \frac{1}{N}\sum_{i=1}^{N} Q_i \tag{2}$$

Perturbation quality measures the difference between the generated adversarial example and the original input, usually represented by the L_2 norm and the L_∞ norm. The L_2 norm reflects the overall perturbation strength and is defined as:

$$\left\|\delta\right\|_2 = \sqrt{\sum_{i=1}^{n}\left(x_i' - x_i\right)^2} \tag{3}$$

whereas the L_∞ norm limits the maximum change applied to any individual element:

$$\|\delta\|_\infty = \max|x_i' - x_i| \tag{4}$$

By remaining nearly imperceptible to the human eye, smaller perturbations enhance stealthiness. To this end, many adversarial training methods, such as Projected Gradient Descent (PGD), enforce constraints like $L_\infty \leq \frac{8}{255}$ to prevent the generation of noticeable noise [10]. Typical adversarial examples illustrating such subtle perturbations are presented in Figure 3, showing how deep models can be misled to incorrect predictions by minimal pixel changes [8]. Together, these metrics provide a comprehensive framework for evaluating black-box attacks, balancing stealth, efficiency, and effectiveness to guide practical adversarial strategy design.

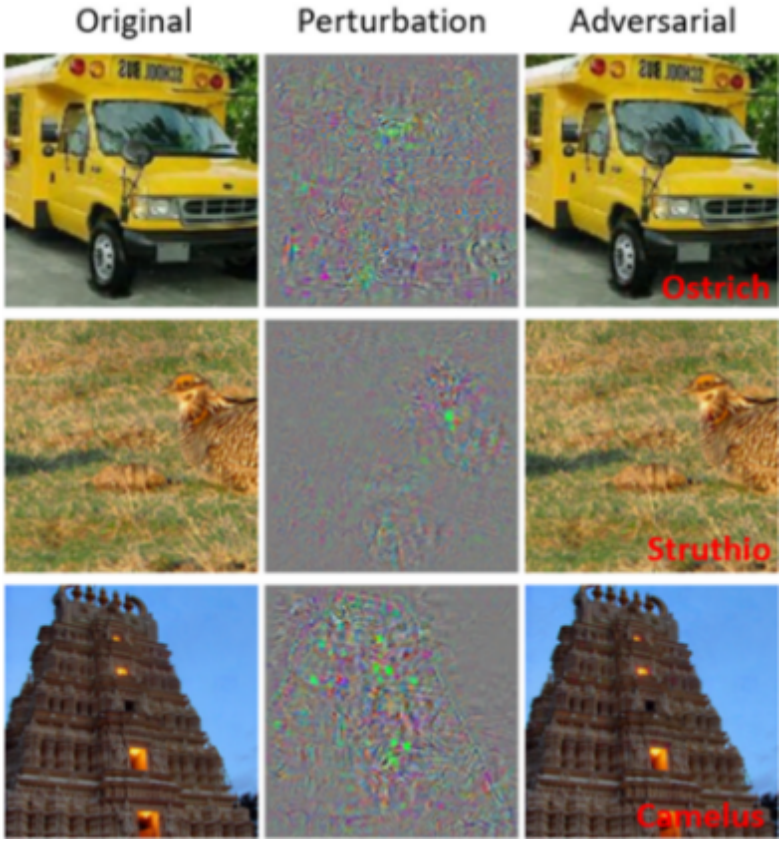


Figure 3. Illustration of adversarial examples

DOI: 10.54254/2755-2721/2025.BJ27247

3.2. Empirical comparison of cross-modal attack effectiveness

The effectiveness of adversarial attacks is shaped by structural characteristics and modality-specific constraints across different data types. In the image domain, the vulnerability stems from models' sensitivity to subtle pixel-level perturbations that can substantially alter learned visual features. These perturbations often remain imperceptible to human observers owing to the insensitivity of the visual system to high-frequency changes, enabling attacks with minimal distortion (e.g., PSNR > 38 dB) [11]. This high stealthiness facilitates adversarial success against architectures such as CNNs and Vision Transformers. In addition, audio and video attacks must satisfy temporal and spectral coherence constraints, adding complexity to the attack process. Specifically, to avoid detection by humans, audio perturbations need to sustain Signal-to-Noise Ratios typically exceeding 20 dB, while video adversarial examples require temporal coherence across frames to preserve naturalness [12]. Such constraints limit the permissible magnitude and structure of perturbations, reducing attack effectiveness compared to images. Text-based models exhibit comparatively higher robustness, attributable to the discrete and semantically rich nature of language data. Minor perturbations, such as character-level edits, often induce grammatical errors or semantic inconsistencies detectable by models [13]. Also, advanced alignment techniques like Reinforcement Learning with Human Feedback (RLHF) have further reinforced robustness by filtering harmful outputs [14]. Consequently, adversarial attacks on text must be carefully designed to preserve semantic coherence while effectively evading detection mechanisms. These modality-specific limits reflect the complex interplay between data structure and human perception, thus emphasizing the need for tailored adversarial strategies that address the unique challenges inherent to each modality.

3.3. Key challenges of efficiency and transferability

The main technical obstacles in black-box adversarial attacks lie in balancing query efficiency with robust transferability across models and tasks. Without access to model internals, black-box attacks depend on iterative query-based strategies like zeroth-order optimization and surrogate modeling to approximate gradients or locate effective perturbations in high-dimensional spaces. These methods narrow the otherwise intractable search space by prioritizing queries that offer the most informative gradients. Typical techniques include adaptive query allocation, importance sampling, and reuse of historical query-response pairs to guide subsequent perturbation updates more efficiently. However, improving query efficiency introduces inherent trade-offs. Excessive query reduction can result in under-explored solution spaces, producing suboptimal or overly conspicuous perturbations that fail to consistently mislead the target model. For example, coarse gradient estimates from sparse queries may misdirect the search trajectory, preventing convergence to minimal yet effective perturbations. In addition, transferability, defined as the ability of adversarial examples generated on one model to generalize to others, is limited by differences in network architectures, training datasets, and learned feature representations. Existing approaches typically enhance transferability only post hoc, using methods such as ensembling or gradient smoothing, rather than integrating transferability objectives directly into the attack generation process. Furthermore, this degradation intensifies when attacking models from different tasks or modalities, due to misaligned features and decision boundaries. More importantly, approaches like meta-learning and domain adaptation boost both query efficiency and transferability, addressing core challenges in black-box adversarial attacks.

DOI: 10.54254/2755-2721/2025.BJ27247

3.4. Real-world constraints and practical considerations

Implementing black-box adversarial attacks in practice is challenged by both technical issues and stringent regulatory, ethical, environmental, and defense-related constraints. In highly regulated sectors like healthcare, finance, and legal systems, strict compliance requirements impose stringent limits on allowable perturbation magnitudes and query frequencies. These constraints significantly limit the practical feasibility of attacks, since high query volumes increase the risk of detection, and large perturbations may breach safety or privacy standards. For example, adversarial manipulations must maintain near-perfect stealth to avoid triggering audit or monitoring systems, which narrows the margin for effective attacks. Environmental variability further complicates real-world attacks, especially in physical scenarios involving adversarial patches or stickers. Factors such as changing lighting conditions, varied viewing angles, and sensor noise can drastically reduce attack robustness and success rates. The ever-changing nature of these conditions requires perturbations to remain effective across diverse physical transformations, posing a major practical challenge. Furthermore, the widespread adoption of defense mechanisms such as input sanitization filters, anomaly detectors, and confidence-based output checks creates an additional barrier. Although these defenses are often straightforward, their effectiveness in real-world systems substantially diminishes attack success, especially when combined with environmental noise. Moreover, fundamental differences in data representation across modalities limit the transferability of attacks. For example, perturbations effective on images rarely apply to text or audio due to distinct features and semantics. This calls for adversarial methods that emphasize adaptability, stability, and control to remain effective under defenses and real-world conditions.

4. Future directions and countermeasures in black-box attack investigation

4.1. Query minimization and meta-learning

The minimization of query counts constitutes a key strategy for enhancing the efficiency and stealth of black-box attacks. Excessive queries not only lead to significant computational overhead but also increase the likelihood of triggering security mechanisms, reducing the practicality and feasibility of the attack. Therefore, recent research has introduced meta-learning and reinforcement learning frameworks that optimize perturbation generation strategies and query paths, enabling effective attacks with minimal model interaction. Meta-learning can learn generalizable perturbation patterns on surrogate models and transfer them to unknown targets, while reinforcement learning leverages reward mechanisms to guide the querying process, allowing the attack to converge to effective perturbations within a limited number of rounds. These techniques greatly reduce reliance on the target model and enhance the efficiency and stability of black-box attacks. Specifically, one method pre-trains transferable universal perturbation vectors across multiple surrogate models, enabling attack success rates above 85% on the CIFAR-10 dataset with only a single-digit number of queries in black-box settings [15]. In addition, another approach constructs a query simulator that integrates diverse querying behaviors via multi-task meta-learning. By aligning the outputs of the surrogate and target models via knowledge distillation, the method reduces the query count for attacking ResNet-50 to about 20, roughly two orders of magnitude lower than conventional approaches [16]. These cases show the practical potential of low-query attack strategies that combine transferability with behavioral fitting mechanisms.

DOI: 10.54254/2755-2721/2025.BJ27247

4.2. Attack generalization and cross-domain transfer

The enhancement of black-box attack transferability across different data types and application domains is critical to achieving generalization and practical deployment. Most existing methods still rely on specific domains or modalities, hence making them difficult to apply to diverse real-world scenarios. To overcome this limitation, feature space decoupling and model structure alignment have been explored to enable cross-modal transfer of attack strategies. On one hand, cross-modal feature extractors trained via contrastive learning can generate unified embedding representations, mitigating semantic discrepancies that often cause attacks to fail. On the other hand, for non-visual modalities, structural modeling at the level of physical mechanisms is required to account for differences in data characteristics and model architectures, ensuring the cross-domain validity of adversarial perturbations. These approaches aim to build attack frameworks with greater robustness and adaptability. In image-based tasks, one study trained a cross-domain feature extractor using contrastive learning, increasing the attack success rate from ImageNet to Oxford Flowers to 86.8%, notably outperforming traditional methods [17]. In non-visual domains, a structurally isomorphic framework combined with a specialized optimizer was proposed for synthetic aperture radar data, achieving a transfer attack success rate of 83.6% even under a 40% domain shift [18]. These results indicate that cross-domain attacks are technically feasible and hold strong potential for deployment in multimodal systems like autonomous driving radars, presenting new challenges for AI security.

4.3. Ethical concerns and practical challenges

As black-box attack techniques continue to advance, their potential risks are becoming increasingly prominent. In high-stakes domains like medical diagnosis and judicial decision-making, adversarial testing may lead to real-world harm, and thus the scope of related research and deployment must be strictly regulated. To ensure research safety and reproducibility, there is an urgent need to establish standardized simulation environments for evaluating attack effectiveness and defensive strategies. Through the emulation of system dynamics, it becomes possible to reconstruct realistic operational scenarios while safeguarding sensitive data, thereby mitigating ethical and privacy concerns during real-world implementation. For highly sensitive AI applications such as facial recognition, ethical review mechanisms should also be implemented to strictly regulate the generation logic and usage boundaries of adversarial examples, preventing misuse and misidentification. For instance, virtual testing platforms for autonomous driving can be constructed to evaluate attack performance without involving real vehicles or users, balancing research progress with data protection. Moreover, expert review has been proposed to avoid misidentifying ordinary users as high-risk individuals, including mistaken entries on terrorism watchlists. Future research on adversarial attacks should strike a balance among technological innovation, engineering practice, and ethical responsibility to ensure that technical advances do not translate into societal harm.

5. Conclusion

This study indicates that limited cross-domain transferability remains a major barrier to the broader applicability of black-box adversarial attacks. In medical imaging tasks, major modality differences lead to marked drops in attack effectiveness. By contrast, text models exhibit higher robustness, largely due to structured syntax and reinforcement learning with human feedback. These findings underscore the challenges of generalizing adversarial examples across diverse real-world scenarios. However, three key limitations remain. Most experiments are conducted on static datasets such as

DOI: 10.54254/2755-2721/2025.BJ27247

CIFAR-10 and ImageNet, lacking evaluation under dynamic, real-world defenses. Besides, the role of architectural synergy during training is poorly understood, limiting transferability gains. Also, in high-risk domains like healthcare, ethical gaps and absent standards obstruct practical deployment. To bridge these gaps, future work should improve efficiency, generalization, and adaptability, and promote shared standards through academic-industry collaboration.

References

[1] Hassan, A., et al. (2024). Navigating IoT security: Insights into architecture, key security features, attacks, current challenges and AI-driven solutions shaping the future of connectivity. Computers, Materials & Continua, 81(3), 3499-3559.
[2] Carlin, N., & Farid, H. (2020). Evading deepfake-image detectors with white- and black-box attacks. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2804-2813.
[3] Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.
[4] Whitley, D. (2001). An overview of evolutionary algorithms: Practical issues and common pitfalls. Information and Software Technology, 43(14), 817-831.
[5] Zhu, Y., et al. (2022). Towards understanding and boosting adversarial transferability from a distribution perspective. IEEE Transactions on Image Processing, 31, 6487-6501.
[6] Tharwat, A., & Schenck, W. (2021). Population initialization techniques for evolutionary algorithms for single-objective constrained optimization problems: Deterministic vs. stochastic techniques. Swarm and Evolutionary Computation, 67, 100952.
[7] Gu, J., et al. (2023). A survey on transferability of adversarial examples across deep neural networks. arXiv: 2310.17626v2 [cs.CV]
[8] Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, 25, 1106-1114.
[9] Wang, J. (2025). The Internet of Things under federated learning: A review of the latest advances and applications. Computers, Materials & Continua, 2025(1), 1-39.
[10] Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, 6, 14410-14430.
[11] Szegedy, C., et al. (2014). Intriguing properties of neural networks. arXiv: 1312.6199v4 [cs.CV].
[12] Wu, X., & Rajan, A. (2021). Blackbox untargeted adversarial testing of automatic speech recognition systems.
[13] Liu, D., et al. (2024). A survey of attacks on large vision-language models: Resources, advances, and future trends.
[14] Liu, A., et al. (2025). PiCo: Jailbreaking Multimodal Large Language Models via \textbf{Pi}ctorial \textbf{Co}de Contextualization. arXiv: 2504.01444v3 [cs.CR]
[15] Fu, J., Sun, J., & Wang, G. (2022). Boosting black-box adversarial attacks with meta learning. Chinese Control Conference (CCC), 7308-7313.
[16] Ma, C., Chen, L., & Yong, J.-H. (2021). Simulating unknown target models for query-efficient black-box attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11835-11844.
[17] Li, Z., et al. (2023). CDTA: A cross-domain transfer-based attack with contrastive learning. In Proceedings of the AAAI Conference on Artificial Intelligence.
[18] Peng, B., et al. (2024). Towards assessing the synthetic-to-measured adversarial vulnerability of SAR ATR. ISPRS Journal of Photogrammetry and Remote Sensing.