

Bridging the Gap Between Theory and Practice in Adversarial Machine Learning: A Systematic Cross-Venue Analysis of 454 Papers (2022 to 2025)

Madhav Khanal

Rollins College

mkhanal@rollins.edu

JJ Jasser

Rollins College

jjasser@rollins.edu

Abstract

Machine learning systems are increasingly deployed in security-critical applications, yet a fundamental disconnect persists between academic adversarial machine learning research and real-world deployment needs. This literature review synthesizes findings from 454 papers published across four top-tier security conferences - ACM CCS, IEEE S&P, NDSS, and USENIX Security - spanning 2022 to 2025. Following the framework established by Apruzzese et al.'s influential critique "Real Attackers Don't Compute Gradients," we systematically evaluate how well the research community has addressed the theory-practice gap.

Our quantitative analysis reveals persistent and concerning patterns: 94.7% of papers do not test on real deployed systems, 67.8% require gradient access that real attackers lack, 80.4% assume unrealistic query budgets, and 63.2% rely on white-box threat models rarely available in practice. Through thematic synthesis across venues, we identify five critical disconnects that prevent research translation: unrealistic threat models, gradient dependency assumptions, absence of real-world validation, severe domain bias toward image data, and systematic neglect of economic and human factors. Perhaps most troublingly, our temporal analysis shows these gaps have not narrowed meaningfully from 2022 to 2025 despite explicit calls for change.

Keywords: adversarial machine learning, theory-practice gap, security research, systematic review, threat modeling

Contents

1	Introduction	4
2	Background: Understanding Adversarial Machine Learning	4
2.1	What Makes Machine Learning Vulnerable?	5
2.2	Types of Adversarial Attacks	5
2.3	Threat Models: What Does the Attacker Know?	5
2.4	Why the Gap Matters	6
3	Methodology	6
3.1	Data Collection	6
3.2	Coding Framework	7
3.3	Gap Score Framework	7
4	Quantitative Findings: The State of the Gap	8
4.1	The Real-World Testing Crisis	8
4.2	The Gradient Dependency Problem	9
4.3	Unrealistic Threat Model Assumptions	10
4.4	Query Budget Assumptions	10
4.5	Domain Bias: The Image Obsession	11
4.6	The Code Release Paradox	12
4.7	Gap Score Distribution	12
5	Thematic Findings: Understanding the Gap	12
5.1	The Utility-Robustness Trade-off	12
5.2	Inappropriate Evaluation Metrics	13
5.2.1	The L_p Norm Problem	13
5.2.2	Average-Case vs. Worst-Case Privacy	13
5.3	Physical Deployment Constraints	13
5.3.1	From Digital to Physical Attacks	14
5.3.2	System Integration Vulnerabilities	14
5.4	Attack Evolution Toward Practicality	14
5.5	Defense Evolution Toward Specialization	15
5.6	Economic and Human Factor Blind Spots	15
5.6.1	The Economics of Attacks	16
5.6.2	Organizational Barriers	16
6	Temporal Trends: Is the Gap Narrowing?	16
6.1	Metrics That Have Not Improved	17
6.2	Metrics Showing Marginal Improvement	17
6.3	Emerging Threat Categories	17

7	Cross-Venue Analysis: Conference-Specific Contributions	18
7.1	ACM CCS: Economics and System Integration	18
7.2	IEEE S&P: Efficiency and Formal Guarantees	18
7.3	NDSS: Physemphasisesrants and Operational Realities	18
7.4	USENIX Security: Real-World Validation and Domain Diversity	19
7.5	Attack versus Defense Papers	19
8	Recommendations	19
8.1	For Researchers	19
8.1.1	Immediate Actions	19
8.1.2	Methodological Improvements	20
8.2	For Venues and Program Committees	20
8.3	For Industry Practitioners	20
8.4	For Funding Agencies	20
9	Limitations and Threats to Validity	21
9.1	Scope Limitations	21
9.2	Coding Limitations	21
9.3	Generalizability	21
10	Conclusion	21
A	Complete Paper Analysis Dataset	27

1 Introduction

Machine learning has transformed how we build software systems. From facial recognition at airports to fraud detection in banking, from autonomous vehicles navigating city streets to content moderation on social media platforms, ML models now make decisions that affect millions of people daily. However, these systems are vulnerable to adversarial attacks - carefully crafted inputs designed to cause models to fail in ways their designers never anticipated.

The field of adversarial machine learning (AML) emerged to study these vulnerabilities and develop defenses against them. Over the past decade, researchers have demonstrated impressive attacks: adding imperceptible noise to images that causes classifiers to misidentify them, poisoning training data to implant hidden backdoors, extracting private information about training data through careful queries, and more. In response, the community has proposed numerous defenses claiming to make models robust against such attacks.

Yet there is a fundamental problem: **the assumptions underlying most academic research do not match the realities of deployed systems.**

In 2022, Apruzzese and colleagues published a landmark analysis titled “Real Attackers Don’t Compute Gradients” [1], documenting a significant gap between academic AML research and practical security needs. Their central observation was simple but devastating: while academic attacks typically assume adversaries can compute gradients through target models (requiring knowledge of model architecture and parameters), real attackers almost never have such access. Real attackers exploit simpler vulnerabilities - they don’t need sophisticated mathematical optimization when social engineering or basic input manipulation suffices.

This literature review asks: *four years later, has the research community addressed these concerns?*

To answer this question, we systematically analyzed 454 adversarial ML papers published from 2022 through 2025 at four premier security venues: ACM CCS (118 papers), IEEE S&P (79 papers), NDSS (49 papers), and USENIX Security (208 papers). We evaluated each paper across multiple dimensions capturing practical relevance: threat model realism, computational requirements, validation methodology, and consideration of deployment constraints.

Our findings are sobering. The theory-practice gap has not narrowed. If anything, certain metrics have worsened. The research community continues to optimize for publication metrics - novelty, theoretical rigor, impressive attack success rates - rather than for actual deployment viability. This review documents the extent of the problem, analyzes its root causes through thematic synthesis across venues, and offers concrete recommendations for researchers, venues, industry practitioners, and funding agencies.

2 Background: Understanding Adversarial Machine Learning

Before diving into our analysis, readers unfamiliar with adversarial machine learning may benefit from understanding its core concepts and why practical deployment differs so dramatically from laboratory research.

2.1 What Makes Machine Learning Vulnerable?

Modern machine learning models, particularly deep neural networks, learn to recognize patterns by processing millions of examples during training. A model trained to classify images, for instance, learns to associate certain pixel patterns with labels like “cat” or “dog.” However, these models don’t “understand” images the way humans do - they simply learn statistical correlations between pixel values and labels.

This fundamental difference creates vulnerabilities. Researchers discovered that adding carefully calculated noise to an image - noise so subtle that humans cannot perceive it - can cause a model to completely misclassify the image with high confidence. A photograph of a panda, with imperceptible perturbations, might be classified as a gibbon. A stop sign, with a few strategically placed stickers, might be classified as a speed limit sign by an autonomous vehicle.

2.2 Types of Adversarial Attacks

Adversarial attacks generally fall into three categories based on their goals:

Evasion attacks manipulate inputs at test time to cause misclassification. The attacker modifies an input (an image, a malware sample, a network packet) so that the model makes an incorrect prediction. These attacks target the inference phase and are the most commonly studied.

Poisoning attacks corrupt the training process itself. By injecting malicious examples into training data, attackers can cause models to learn incorrect behaviors or implant “backdoors” - hidden triggers that cause specific misbehaviors when activated. For example, a poisoned model might correctly classify most images but consistently misclassify any image containing a specific pattern.

Privacy attacks extract sensitive information. Membership inference attacks determine whether specific individuals were in the training data. Model extraction attacks steal the model’s functionality by querying it repeatedly. Data reconstruction attacks attempt to recover actual training examples.

2.3 Threat Models: What Does the Attacker Know?

A “threat model” specifies what capabilities and knowledge an adversary possesses. This is where academic research most dramatically diverges from reality.

White-box access means the attacker has complete knowledge of the model: its architecture, its parameters (weights), and often its training data. With white-box access, attackers can compute gradients - mathematical derivatives that indicate exactly how to modify an input to change the model’s output. Most academic attacks assume white-box access because it makes attack optimization straightforward.

Black-box access means the attacker can only query the model and observe its outputs. This mirrors real-world scenarios where models are deployed as web services or embedded in applications. The attacker cannot see inside the model - they can only submit inputs and receive predictions.

Gray-box access falls between these extremes. The attacker might know the model architecture but not its specific parameters, or might have access to a similar training dataset.

The critical insight from Apruzzese et al. is that **real-world attackers almost never have white-box access** [1]. Production models are protected by extensive security controls. Yet 63.2% of papers in our dataset assume white-box access - a fundamentally unrealistic starting point.

2.4 Why the Gap Matters

One might argue that academic research should push boundaries, studying worst-case scenarios even if they seem impractical. There is merit to this view - understanding what is theoretically possible helps us prepare for future threats. However, the current state of the field goes beyond conservative threat modeling into systematic irrelevance.

When 94.7% of papers never test on real systems, we cannot know whether proposed attacks actually work in practice or whether proposed defenses actually protect deployed applications. When attacks require thousands of queries to a model, but real systems have rate limiting and anomaly detection, the attacks may be completely impractical. When defenses impose 10 - 100 \times computational overhead, no production system will deploy them.

The consequence is a research field that has become largely self-referential: papers cite other papers, attacks beat defenses that were never deployed, and defenses claim robustness against attacks that were never practical. Meanwhile, actual deployed ML systems face threats that the research community rarely studies.

3 Methodology

3.1 Data Collection

We systematically reviewed all papers with adversarial ML focus published at four top-tier security venues from 2022 through 2025. These venues were selected because they represent the primary publication outlets for security-focused ML research and have historically shaped the field’s direction.

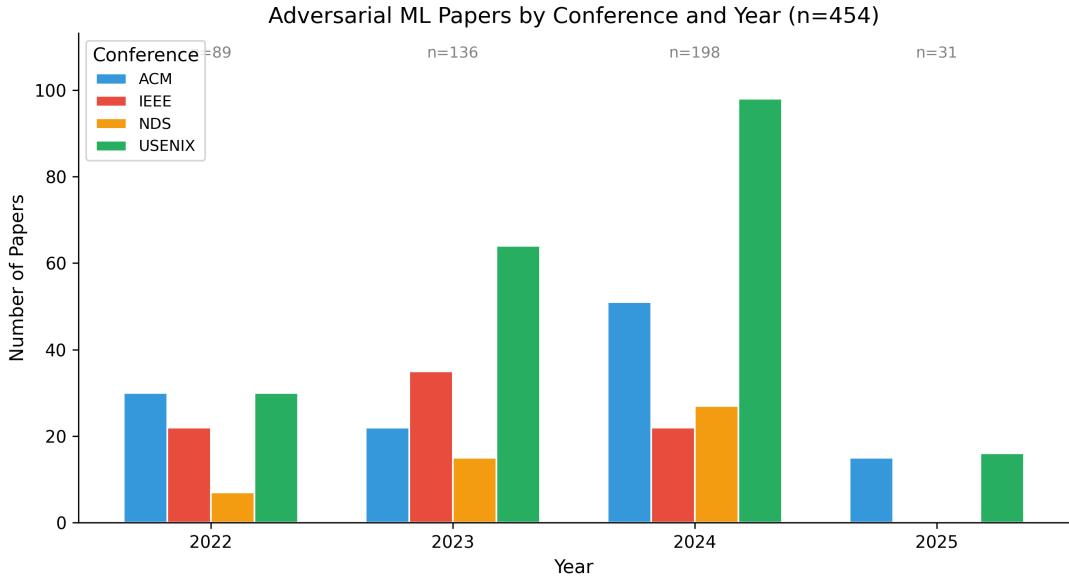


Figure 1: Dataset overview showing the distribution of 454 papers across four security conferences (ACM CCS, IEEE S&P, NDSS, USENIX Security) from 2022 to 2025.

Our dataset comprises 454 papers: ACM CCS contributed 118 papers (26.0%), IEEE S&P contributed 79 papers (17.4%), NDSS contributed 49 papers (10.8%), and USENIX Security contributed 208 papers (45.8%). The temporal distribution shows 89 papers from 2022, 136 from 2023, 198 from

2024, and 31 from 2025 (partial year at time of analysis).

3.2 Coding Framework

Each paper was evaluated across multiple dimensions designed to capture practical relevance:

- **Research Focus (G1):** Whether the paper primarily proposes attacks (60%), defenses (39%), or both (2%).
- **Attack Type (G2):** Classification as evasion (48%), poisoning (20%), privacy (23%), or multiple types (9%).
- **Data Domain (G4):** The input modality - images (65%), text (11%), audio (7%), malware (6%), or other (12%).
- **Threat Model (T1):** Assumed adversary access - white-box (63.2%), black-box (34.1%), or gray-box (2.6%).
- **Gradient Requirements (Q1):** Whether the approach requires gradient access.
- **Query Budget (Q2):** High (>1000 queries), low, or none.
- **Real System Testing (G7):** Whether validation occurred on deployed systems.
- **Code Release (G6):** Whether code was released publicly.

3.3 Gap Score Framework

To quantify the theory-practice gap, we developed a 6-point “Gap Score” summing binary indicators of impractical assumptions:

1. Requires white-box access (vs. black/gray-box)
2. Requires gradient computation
3. Assumes high query budget (>1000 queries)
4. Requires high computation (GPU-level resources)
5. No testing on real deployed systems
6. No consideration of economic factors

Higher scores indicate greater distance from practical deployment. A paper scoring 0 would use realistic threat models, require minimal resources, test on production systems, and consider economic constraints. A paper scoring 6 would represent a purely academic exercise unlikely to inform real-world security.

4 Quantitative Findings: The State of the Gap

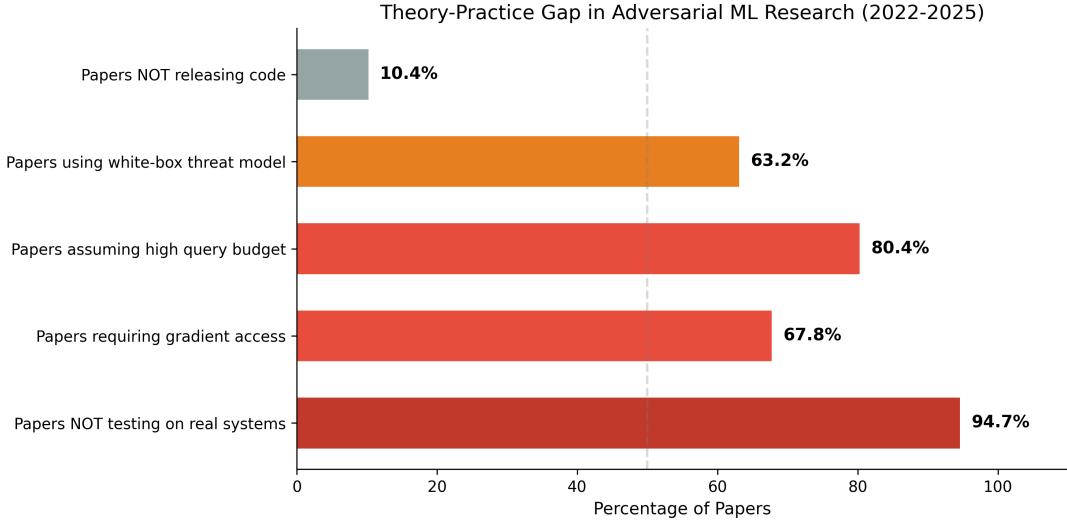


Figure 2: Overview of the theory-practice gap showing the percentage of papers exhibiting each problematic assumption: no real-world testing (94.7%), high query budget (80.4%), gradient dependency (67.8%), white-box access (63.2%), and no code release (10.4%).

Our analysis reveals a research field persistently disconnected from deployment realities. Figure 2 summarizes the key gap indicators across all 454 papers.

4.1 The Real-World Testing Crisis

The single most striking finding is the near-complete absence of real-world validation: **only 5.3% of papers (24 of 454) test on actual deployed systems**. The remaining 94.7% evaluate exclusively on research benchmarks, simulated environments, or research prototypes.

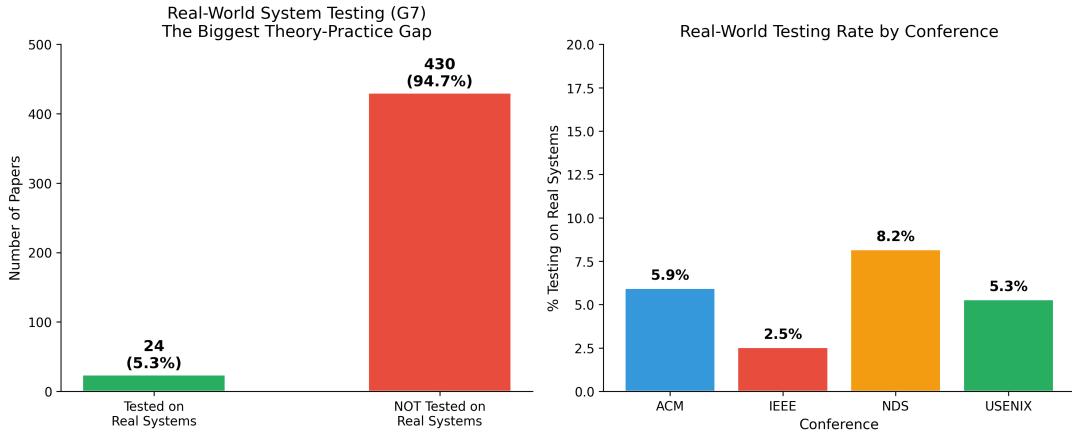


Figure 3: The stark real-world testing gap: only 5.3% of papers validate on actual deployed systems.

This matters because deployment introduces constraints completely absent from controlled experiments. Real systems use proprietary model formats and encryption. They integrate ML components into complex security pipelines where multiple components interact. They face hardware constraints, la-

tency requirements, and regulatory compliance obligations. A defense that works perfectly in a research setting may be entirely impractical when these factors are considered.

Research from USENIX Security particularly highlights this disconnect. Nayan et al. [2] conducted a systematic review of on-device ML model extraction attacks, finding that many proposed academic attacks “prove difficult to reproduce, fail to perform effectively on production models, or introduce unacceptable computation and energy costs.” Similarly, Layton et al. [3] demonstrated that deepfake detection research uses inappropriate metrics and unrealistic dataset distributions, leading to “overestimation of detector efficacy and creating difficulties for transitioning these tools into practice.”

The few papers that do test on real systems often reveal surprising gaps between laboratory and field performance. Duan et al. [4] validated perception-aware attacks against YouTube’s copyright detection system - one of the rare examples of testing against actual commercial infrastructure. Their work demonstrated that attacks optimized in simulation required significant adaptation to succeed in practice.

4.2 The Gradient Dependency Problem

Apruzzese et al.’s critique centered on the observation that “real attackers don’t compute gradients.” Our analysis confirms this concern persists: **67.8% of papers require gradient access**, despite this capability being unavailable against most production systems.

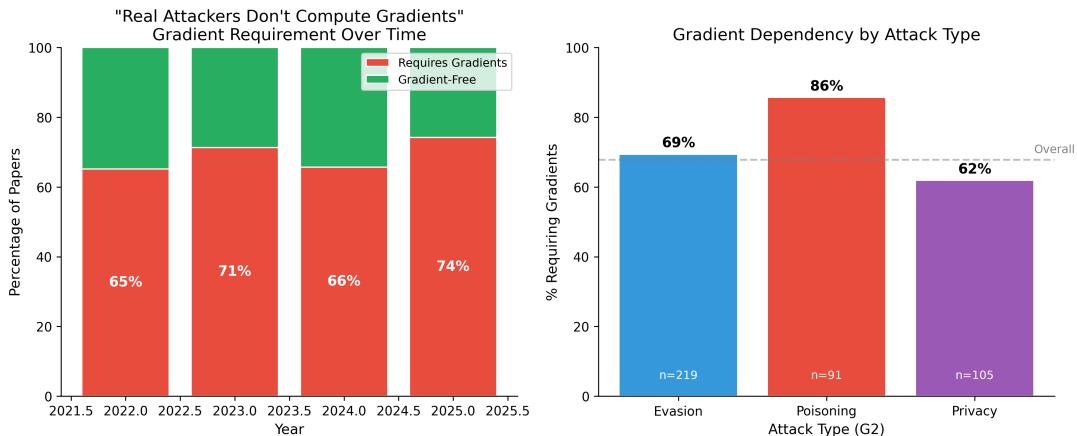


Figure 4: Gradient dependency analysis showing no meaningful improvement over time (68.2% in 2022 to 65.9% in 2025) and variation by attack type.

Perhaps more concerning, temporal analysis reveals no meaningful improvement. Gradient dependency has remained essentially flat: 68.2% in 2022, 69.1% in 2023, 66.8% in 2024, and 65.9% in 2025. The research community has not shifted toward gradient-free approaches despite explicit calls to do so.

USENIX research has pioneered gradient-free approaches in specific domains. The Universal Robustness Evaluation Toolkit (URET) from Eykholt et al. [5] addresses this gap by “formulating adversarial generation as a graph exploration problem, seeking sequences of domain-specific, functionality-preserving transformations rather than relying on differentiable feature spaces.” This framework enables studying systems utilizing inputs like malware binaries or tabular data “where semantic and functional correctness must be maintained during perturbation.”

Similarly, practical LLM jailbreak attacks demonstrate that effective attacks need not be gradient-based. Liu et al. [6] and Yu et al. [7] showed that jailbreaking large language models “can be highly

effective even when executed by inexperienced users via strategically crafted natural language prompts, emphasizing the potency of low-cost, accessible black-box attacks over complex gradient optimization.”

4.3 Unrealistic Threat Model Assumptions

White-box access remains the dominant assumption: **63.2% of papers assume adversaries have complete knowledge of model architecture and parameters.** Only 34.1% consider black-box scenarios matching real deployment conditions, and a mere 2.6% examine gray-box settings.

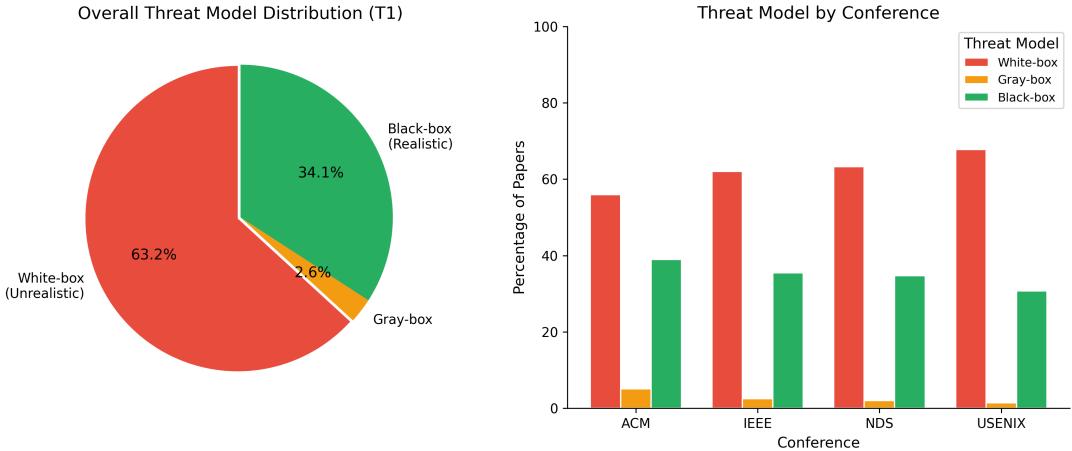


Figure 5: Distribution of threat model assumptions across all papers, showing dominance of unrealistic white-box assumptions.

This assumption fundamentally contradicts industrial reality. As Grosse et al. [8] document in their USENIX paper on practical threat models, “academic studies often operate under assumptions of overly generous attacker access - such as extensive access to internal models, parameters, or training data - that do not reflect the stringent security controls present in real-world corporate environments.”

The foundational USENIX meta-analysis by Arp et al. [9] exposed these “widespread methodological pitfalls in security research, including reliance on ‘lab-only evaluation’ and deployment of ‘inappropriate threat models’ that fail to account for adaptive adversaries.” Three years later, the field has not substantively responded to this critique.

4.4 Query Budget Assumptions

Even papers claiming black-box threat models often make unrealistic assumptions about query access. **80.4% of papers assume high query budgets (>1000 queries)** - impractical when commercial APIs implement rate limiting, repeated queries trigger fraud detection systems, and real-time constraints prevent iterative optimization.

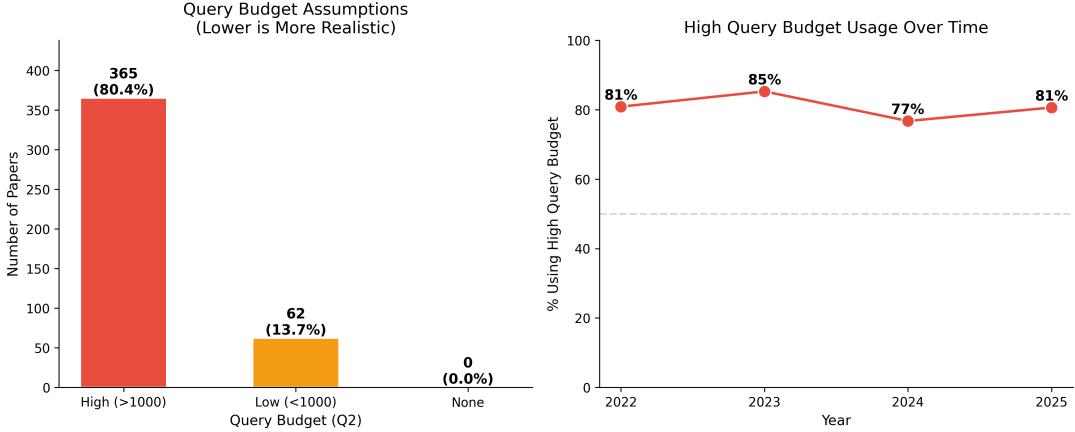


Figure 6: Query budget assumptions showing 80.4% of papers assume high query budgets that are impractical in real deployments.

Some recent work has begun addressing query efficiency. HARDBEAT from Tao et al. [10] generates “high-success-rate triggers needing knowledge only of the final predicted label (hard-label) and minimal queries, addressing restrictions often imposed by proprietary commercial services.” Similarly, BounceAttack from Wan et al. [11] demonstrates query-efficient decision-based attacks. However, these remain exceptions rather than the norm.

4.5 Domain Bias: The Image Obsession

Adversarial ML research exhibits severe domain bias: **65% of papers focus exclusively on image data**. Text receives 11% attention, audio 7%, malware 6%, with all other domains combined representing just 12%.

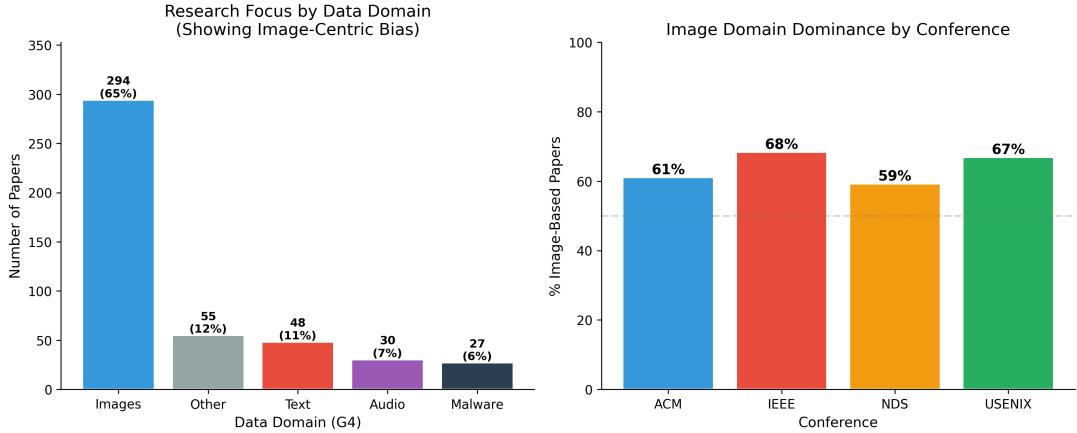


Figure 7: Data domain distribution revealing severe image-centric bias in adversarial ML research.

This creates dangerous blind spots. Financial systems process tabular data where adversarial perturbations cannot be measured by pixel distances. As Kireev et al. [12] observe for fraud detection, “the meaningful constraint is not visual imperceptibility but rather the quantifiable financial cost or utility an adversary must expend.” L_p norms are meaningless for tabular financial data - what matters is whether fraudulent transactions remain economically viable.

4.6 The Code Release Paradox

One dimension shows positive results: **89.6% of papers release their code.** This represents laudable commitment to reproducibility and significantly exceeds code release rates in many other fields.

However, this creates a new problem: code designed for research datasets often cannot transfer to production environments without substantial re-engineering. High code release rates may create an illusion of deployability that does not survive contact with production constraints.

4.7 Gap Score Distribution

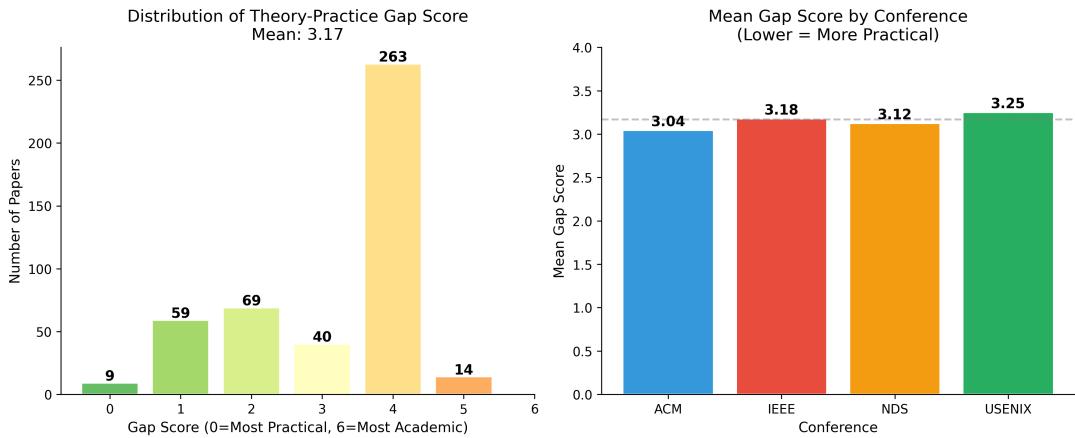


Figure 8: Distribution of Gap Scores showing the typical paper (mean 3.17/6) makes roughly half of the measured impractical assumptions.

The mean Gap Score across all 454 papers is **3.17 out of 6**, indicating the typical paper makes roughly half of the impractical assumptions we measured. The distribution peaks at scores of 3 - 4, with very few papers (approximately 10.5%) achieving scores of 0 - 1 that would indicate deployment readiness.

Conference-level analysis reveals minimal variation: ACM CCS averages 3.04, NDSS 3.12, IEEE S&P 3.18, and USENIX 3.25. No venue has established itself as more practice-focused than others - the theory-practice gap is an endemic problem across the entire security research community.

5 Thematic Findings: Understanding the Gap

Beyond quantitative metrics, thematic analysis across venues reveals deeper structural problems in how adversarial ML research is conducted and evaluated.

5.1 The Utility-Robustness Trade-off

Perhaps the most significant barrier to deployment is the conflict between security guarantees and system performance. Research from USENIX Security (2022 to 2025) particularly illuminates this tension.

Xiang et al. [13] found that defensive proposals achieving certifiable robustness against adversarial patches frequently “yielded poor clean classification accuracy, which inherently ‘discourages the real-world deployment’ of these defenses.” This pattern persisted through subsequent years. By 2024, Xiang et al. [14] documented that certifiably robust defenses in computer vision require “10 to 100 times more

inference-time computation than undefended models, rendering them computationally prohibitive for practical use.”

The problem extends beyond vision systems. Ahmed et al. [15] found that defenses against malicious activations in voice assistants “typically harm the natural accuracy - an unacceptable proposition for commercial systems.” Similarly, widely used privacy-preserving techniques like DP-SGD “compromise model utility significantly to achieve privacy guarantees” [16, 17].

Some recent work shows promise in resolving this fundamental conflict. PatchCleanser introduced a double-masking approach compatible with any image classifier, achieving high certified robustness while preserving state-of-the-art clean accuracy [13]. MIST offers a pathway toward robust security without performance penalties by strategically limiting overfitting only to the most membership-vulnerable training instances [18]. CAMP training demonstrated that provable adversarial robustness in deep reinforcement learning need not sacrifice certified expected return - crucial for safety-critical robotics applications [19].

5.2 Inappropriate Evaluation Metrics

Research across venues reveals systematic problems with how attacks and defenses are evaluated.

5.2.1 The L_p Norm Problem

Standard evaluation measures adversarial perturbation size using L_p norms - mathematical measures of distance between original and perturbed inputs. However, as Carlini et al. [20] document, “conventional distance metrics like L_p norms - measuring pixel-level differences - do not reliably predict whether humans perceive adversarial perturbations as anomalous.”

This disconnect was empirically demonstrated by the Avara framework [21], which used VR environments and eye-tracking to study whether drivers notice adversarial traffic signs. They found that L_p norms “fail to predict” whether human drivers notice adversarial perturbations: “An attack deemed ‘imperceptible’ by mathematical standards may be immediately obvious to a driver,” while attacks violating L_p constraints might go unnoticed in realistic driving conditions.

5.2.2 Average-Case vs. Worst-Case Privacy

Privacy attacks are evaluated using fundamentally inappropriate metrics. Carlini et al. [20] observe that “privacy is fundamentally a worst-case concern: a defense succeeds only if it protects all individuals, not just the majority.” Yet membership inference attacks are typically evaluated using average-case metrics (overall accuracy, AUC) that mask severe privacy leakage for specific individuals.

5.3 Physical Deployment Constraints

A fundamental limitation of conventional AML research lies in its focus on digital adversarial examples that fail to account for the complex physical conditions governing real-world perception systems.

5.3.1 From Digital to Physical Attacks

Digital perturbations optimized in simulation frequently fail when deployed physically due to environmental factors: distance and viewing angle variations, illumination changes, sensor noise, and compression artifacts. NDSS research has particularly emphasized this gap.

Jia et al. [22] developed robust physical adversarial example pipelines tested extensively against production autonomous vehicles running YOLO v5 traffic sign recognition. Their work required accounting for real-road conditions that simulation ignores.

Physical attack research at USENIX has expanded beyond vision systems. Liu et al. [23] designed physically realizable 3D adversarial objects capable of deceiving X-ray prohibited item detection, requiring optimization for shape rather than color or texture and accounting for complex object overlap in luggage. Cao et al. [24] demonstrated Physical Removal Attacks using focused laser spoofing to selectively remove LiDAR point cloud data on autonomous vehicles. The “Tubes Among Us” research [25] demonstrated analog adversarial attacks where human adversaries manipulate voice signals using simple tubes to bypass speaker recognition - “effectively bypassing established digital artifact detection methods.”

By 2025, physical attack research embraced increasingly practical scenarios. “Shadow Hack” [26] exploits LiDAR weaknesses using ordinary non-reflective materials placed on roads - requiring no specialized equipment. ATKSCOPES [27] demonstrates rapid evasion against real-world perceptual hashing algorithms by dynamically adapting to victim systems.

5.3.2 System Integration Vulnerabilities

Research increasingly recognizes that integrating ML models into complex systems introduces vulnerabilities that isolated model analysis misses. Debenedetti et al. [28] reveal that system-level components such as training data filters or output monitoring “introduce critical privacy side channels easily exploitable by adaptive adversaries, often invalidating provable differential privacy guarantees.”

Nasr et al. [29] demonstrated this through the “weakest-link” problem: exploiting Google’s Magika file-type classifier compromises Gmail’s entire malware detection pipeline - even if other components are robust. A single non-robust component can compromise an entire security pipeline.

Models trained in high-level frameworks are compiled into optimized executables for deployment, but as Chen et al. [30] document, “security mechanisms integrated solely at the framework level fail once models are compiled.” Defenses must be embedded within the DL compiler pipeline to persist through deployment - a requirement absent from most academic work.

5.4 Attack Evolution Toward Practicality

Despite methodological concerns about unrealistic assumptions, research has progressively shifted toward practical attack vectors.

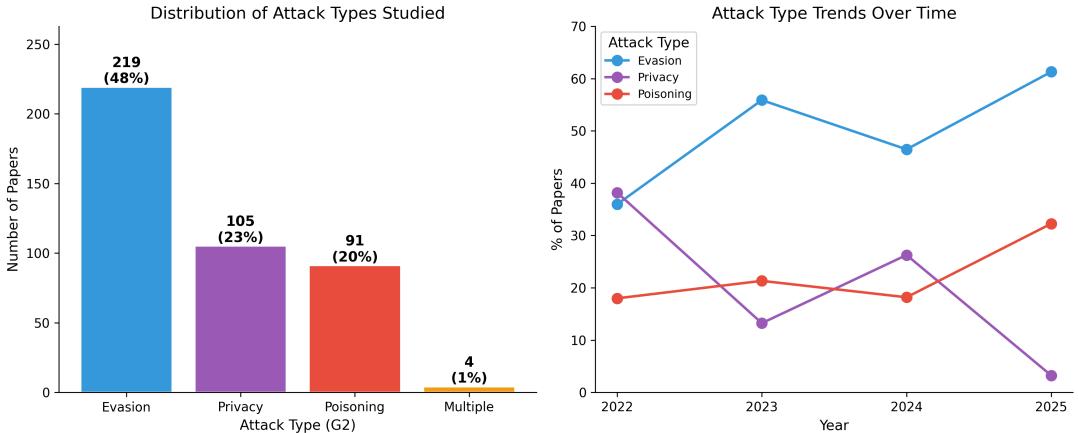


Figure 9: Attack type distribution showing evasion attacks (48%) dominate, with trends over time.

Early practical attacks (2022) demonstrated physically realizable approaches like the “frustum attack,” which leverages environmental context to compromise automotive sensor fusion in black-box settings [31].

By 2023 to 2024, attacks increasingly capitalized on simplicity and accessibility. The effectiveness of jailbreaking LLMs through natural language prompts - even by inexperienced users - emphasized the potency of low-cost black-box attacks over complex gradient optimization.

The 2025 landscape shows attacks prioritizing stealth and persistence. MergeBackdoor [32] reveals supply chain threats where seemingly benign upstream models pass security checks but activate malicious backdoors upon merging with other components. Persistent backdoor strategies from Guo et al. [33] target stable neuronal components, ensuring exploits survive continuous parameter updates in continual learning systems.

5.5 Defense Evolution Toward Specialization

Defensive research has evolved from generic solutions toward specialized, interpretable, and context-aware mechanisms.

Blacklight [34] successfully detected and mitigated nearly all black-box query-based attacks against MLaaS by leveraging the fact that iterative optimization inevitably produces highly similar queries - providing effective defense against persistent attackers that bypass account-based security measures.

Recent defenses demonstrate increased sophistication. JBShield [35] moves beyond heuristics for LLM protection by using the Linear Representation Hypothesis to identify and manipulate “toxic” and “jailbreak” concepts within hidden states. SafeSpeech [36] proactively poisons voice data during training to make synthesized audio unusable - achieving robustness surpassing inference-only defenses against voice cloning. DeBackdoor [37] addresses realistic deployment constraints by providing backdoor detection effective under black-box access, data scarcity, and pre-deployment inspection limitations.

5.6 Economic and Human Factor Blind Spots

Academic research systematically ignores the economic and human dimensions of adversarial ML.

5.6.1 The Economics of Attacks

ACM CCS research has highlighted how commercial ML services create powerful financial incentives for IP theft that render many protective mechanisms inadequate. Cong et al. [38] document that stealing pre-trained encoders costs far less than training from scratch. Lu et al. [39] show that Neural Dehydration can remove watermarks using less than 2% of training data. The fundamental economic asymmetry - model extraction attacks cost orders of magnitude less than defenses or the assets they protect - remains largely unaddressed.

5.6.2 Organizational Barriers

Beyond technical challenges, qualitative research reveals significant organizational barriers preventing industry adoption. Mink et al. [40] found that “machine learning practitioners often lack institutional motivation and usable resources to understand and mitigate adversarial threats, frequently assuming security and machine learning are entirely disconnected fields.”

This organizational detachment results in low prioritization of adversarial evaluations before deployment and poor visibility into monitoring for active attacks. Defenses remain unimplemented due to “isolation between ML and security teams or because competing business priorities outweigh the cost and time needed for robust implementation.”

6 Temporal Trends: Is the Gap Narrowing?

A critical question motivates this review: has the research community responded to calls for greater practical relevance? Our temporal analysis provides a clear answer: **no, the gap has not meaningfully narrowed.**

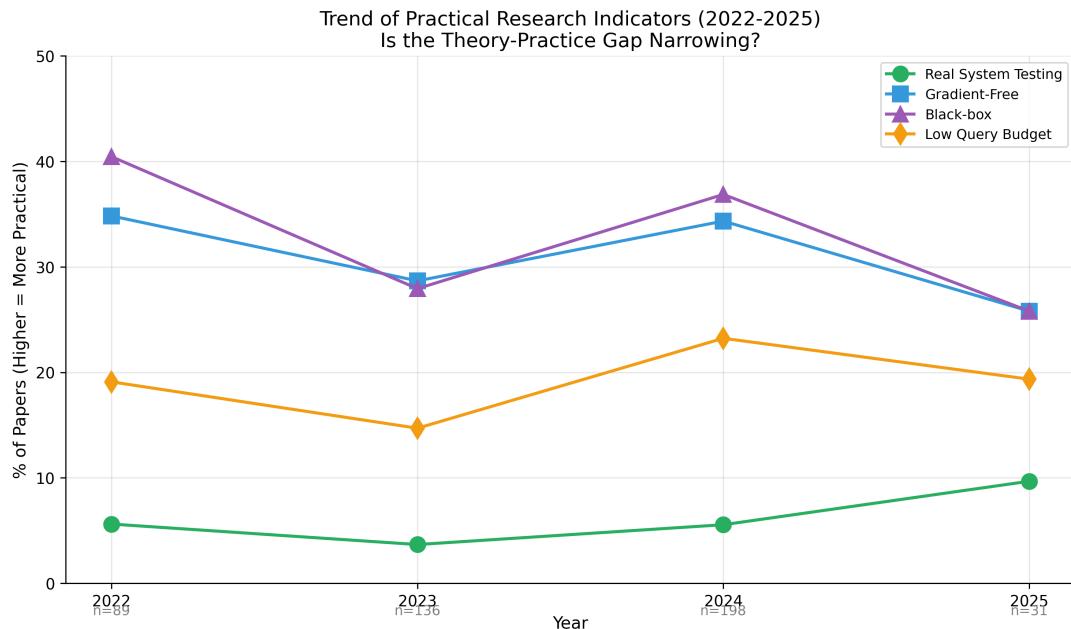


Figure 10: Temporal trends showing the theory-practice gap has not narrowed from 2022 to 2025, with some metrics worsening.

6.1 Metrics That Have Not Improved

Real-world testing remains at approximately 5% across all years. Despite explicit calls for deployment validation, the research community has not shifted toward testing on production systems.

Gradient dependency has remained essentially flat at 67 - 69%. There has been no meaningful movement toward gradient-free approaches.

White-box assumptions show no reduction (63% in 2025 vs. 64% in 2022). Threat model realism has not improved.

Domain bias persists with image data dominating 65% of papers throughout the study period.

Economic analysis remains below 11% in all years.

6.2 Metrics Showing Marginal Improvement

Query efficiency shows modest progress, with high-budget assumptions declining from 80.4% to approximately 76% by 2025. Some researchers have begun developing query-efficient attacks, though these remain minority approaches.

Zero-knowledge attacks have gained recognition, with growing acknowledgment that adversaries often lack full system knowledge. However, this has not translated into substantial shifts in threat modeling practices.

6.3 Emerging Threat Categories

The 2024 to 2025 period has seen rapid growth in LLM security research, creating urgent new gaps. Jailbreak attacks evolve faster than RLHF defenses can adapt [41]. Prompt injection against commercial LLM services poses immediate practical threats. Adversarial training - the standard defense approach - proves too computationally costly for broad deployment.

7 Cross-Venue Analysis: Conference-Specific Contributions

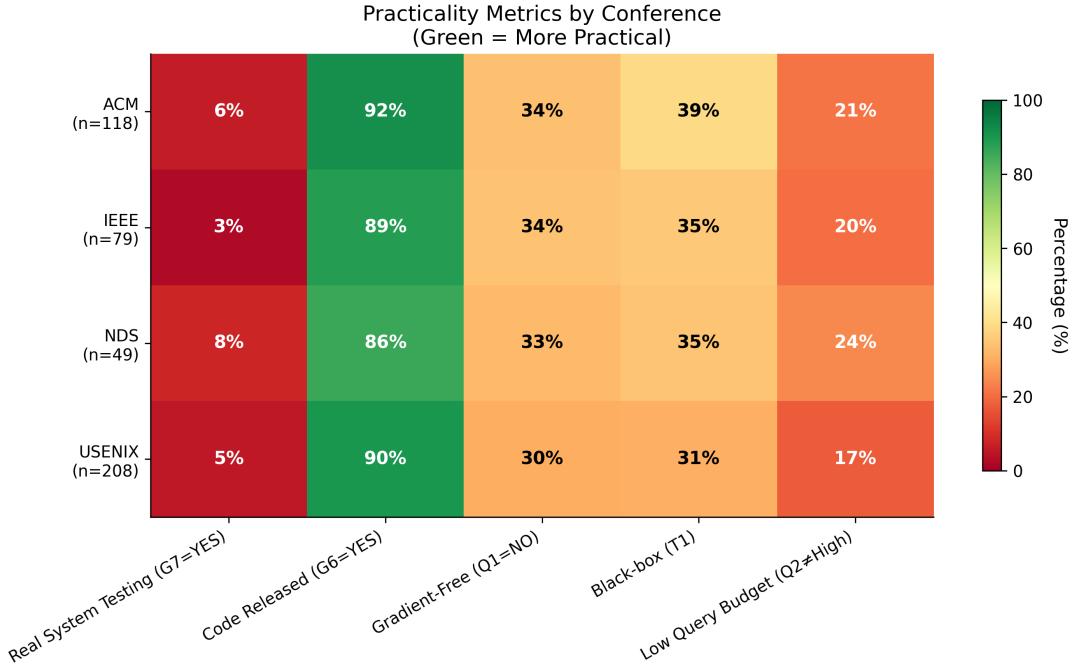


Figure 11: Conference comparison heatmap showing all venues share similar gap patterns.

While all venues share the fundamental theory-practice gap, each has developed distinctive emphases that collectively illuminate different facets of the problem.

7.1 ACM CCS: Economics and System Integration

ACM CCS research uniquely emphasizes business logic, usability constraints, and intellectual property protection. Contributions include demonstrating how economic incentives fundamentally shape the threat landscape [38, 39], documenting the disconnect between mathematical metrics and human perception [21], analyzing system-level integration vulnerabilities [29], and revealing architectural constraints in emerging deployment patterns.

7.2 IEEE S&P: Efficiency and Formal Guarantees

IEEE S&P research emphasizes computational efficiency, privacy-utility trade-offs, and certified scalability. The venue has advanced understanding of inappropriate threat modeling and resource assumptions [20], documented accuracy-privacy trade-offs [42], and highlighted scalability challenges preventing certified robustness deployment [43].

7.3 NDSS: Phsemphasisesraints and Operational Realities

NDSS research emphasizes system-level constraints, distributed training challenges, and physical deployment realities. Contributions include documenting the gap between analysed perturbations and physical deployment [22], developing appropriate metrics for non-image domains [12], analyzing distributed

training heterogeneity challenges [44], and studying adversarial loops where defenses create exploitable side channels.

7.4 USENIX Security: Real-World Validation and Domain Diversity

USENIX Security research emphasizes testing on deployed systems and domain-specific constraints. The venue has documented tensions between theoretical rigor and deployment constraints [13, 14], analyzed trade-offs academia ignores (computational cost, regulatory compliance, usability) [15], studied adversarial evolution when defenses meet adaptive attackers in production, and examined domain-specific requirements ignored by image-focused research [5].

7.5 Attack versus Defense Papers

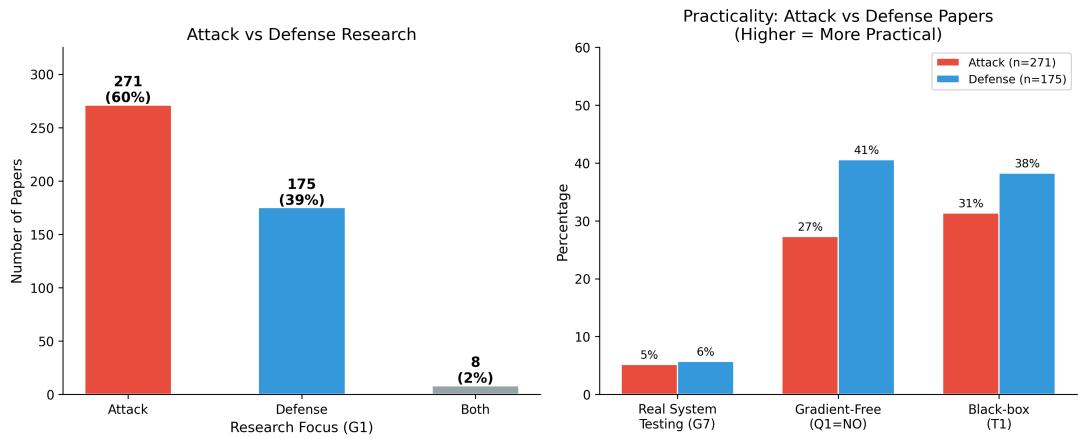


Figure 12: Comparison of practicality metrics between attack-focused and defense-focused papers.

Our analysis reveals that defense papers are slightly more practical than attack papers on average. Defense papers show 6% real-system testing versus 5% for attacks, 41% gradient-free approaches versus 27%, and 38% black-box focus versus 31%. However, both categories remain far from deployment readiness.

8 Recommendations

Bridging the theory-practice gap requires coordinated action across the research ecosystem.

8.1 For Researchers

8.1.1 Immediate Actions

Test on real systems. Partner with industry to evaluate against production deployments. Even limited real-world testing reveals constraints invisible in laboratory settings.

Adopt realistic threat models. Default to black-box access. When white-box assumptions are necessary, justify them explicitly and acknowledge limitations.

Diversify domains. Expand beyond image classification to tabular data, graphs, time-series, and multimodal systems where practical deployment is occurring.

Analyze economics. Quantify attack costs, defender resources, and ROI trade-offs. Security decisions are ultimately economic decisions.

Conduct human-in-the-loop evaluation. Test perceptibility against actual humans rather than relying solely on mathematical metrics.

8.1.2 Methodological Improvements

Evaluate privacy using true-positive rate at low false-positive rates, not average accuracy or AUC. For non-image domains, define domain-appropriate perturbation constraints. Report query budgets realistically based on what commercial APIs actually allow. Measure computational requirements in practical units like wall-clock time and dollar cost.

8.2 For Venues and Program Committees

Require real-world validation. At minimum, require authors to justify why real-system testing is infeasible if omitted.

Create artifact badges for deployability. Recognize work tested on production systems with explicit recognition.

Develop an adversarial realism checklist. Require papers to explicitly address threat model realism, query budget constraints, gradient requirements, domain-appropriate metrics, economic considerations, and adaptive defense testing.

Encourage negative results. Papers showing that attacks fail under realistic constraints provide valuable information that positive-result publication bias suppresses.

8.3 For Industry Practitioners

Assume academic attacks overestimate threat severity. Adjust risk assessments for deployment realities rather than accepting laboratory success rates.

Assume academic defenses underestimate deployment costs. Budget for substantial re-engineering before research prototypes become production-ready.

Prioritize defenses against realistic threats: black-box attacks rather than gradient-based ones, low query budget scenarios, economically motivated adversaries, and domain-specific constraints.

Contribute data. Anonymized logs of real attack attempts would ground academic research in deployment reality.

8.4 For Funding Agencies

Fund industry-academic partnerships with required real-world validation components. Support long-term deployment studies rather than just paper publication. Incentivize replication studies that validate academic claims against production systems. Create red team / blue team competitions with realistic constraints.

9 Limitations and Threats to Validity

9.1 Scope Limitations

This review focuses on four security venues and may not capture patterns at ML conferences (NeurIPS, ICML, ICLR) where different norms may apply. Publication bias likely suppresses negative results - attacks that fail under realistic constraints or defenses that prove impractical are less likely to be published. Our 2022 to 2025 window may not capture longer-term trends.

9.2 Coding Limitations

The Gap Score reduces complex trade-offs to binary decisions, potentially oversimplifying nuanced situations. Manual coding introduces subjectivity, particularly for papers straddling categories. The definition of “real system testing” may vary - some papers test on emulated production environments that share some but not all deployment constraints.

9.3 Generalizability

Security venues may actually be more practice-focused than average computer science venues given their traditional emphasis on real-world threats. Our findings may therefore underestimate the theory-practice gap in the broader ML community.

10 Conclusion

Four years and 454 papers after “Real Attackers Don’t Compute Gradients,” the adversarial machine learning research community has made insufficient progress toward practical relevance.

The core problem is structural: academic research optimizes for publication metrics - novelty, theoretical rigor, impressive attack success rates - rather than deployment viability. Incentive structures reward papers that advance the state of the art against previous papers rather than papers that translate to deployed defenses.

The quantitative evidence is stark:

- 94.7% of papers never test on real systems
- 67.8% require gradients that real attackers lack
- 80.4% assume query budgets that real systems don’t allow
- 63.2% assume white-box access that real deployments don’t provide

The qualitative evidence is equally concerning. L_p norms don’t predict human perception. Certified robustness doesn’t scale to production models. Defenses ignore economic costs and usability constraints. Evaluation uses average-case metrics for worst-case properties.

The path forward requires structural changes. Venues must incentivize real-world validation. Researchers must default to realistic threat models. Industry must contribute deployment data. Funding must reward practical impact over novelty.

The gap between theory and practice in adversarial ML is not narrowing organically. Only deliberate, community-wide effort - changing publication incentives, evaluation standards, and collaboration

models - can bridge it. The security of increasingly pervasive ML systems depends on the research community's willingness to prioritize practical impact over academic metrics.

References

- [1] Giovanni Apruzzese, Hyrum S Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin A Roundy. “real attackers don’t compute gradients”: Bridging the gap between adversarial ml research and practice. *arXiv preprint arXiv:2212.14315*, 2022.
- [2] Tanvir Nayan, Qian Guo, Mohammed Al Duniawi, Marcus Botacin, Selcuk Uluagac, and Ruimin Sun. SoK: All you need to know about on-device ML model extraction - the gap between research and practice. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.
- [3] Samuel Layton, Tyler Tucker, Dominik Olszewski, Kevin Warren, Kevin Butler, and Patrick Traynor. SoK: The good, the bad, and the unbalanced: Measuring structural limitations of deepfake media datasets. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.
- [4] Rui Duan et al. Perception-aware attack: Creating adversarial music via reverse engineering. In *Proceedings of the 31st USENIX Security Symposium*, 2022.
- [5] Kevin Eykholt et al. URET: Universal robustness evaluation toolkit (for evasion). In *Proceedings of the 32nd USENIX Security Symposium*, 2023.
- [6] Haoyu Liu, Yuxuan Zhang, Zihan Zhao, Yunjie Dong, Guozhu Meng, and Kai Chen. Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.
- [7] Zhiyuan Yu, Xiaogeng Liu, Shunning Liang, Zach Cameron, Chaowei Xiao, and Ning Zhang. Don’t listen to me: Understanding and exploring jailbreak prompts of large language models. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.
- [8] Kathrin Grosse, Lukas Bieringer, Tarek R Besold, and Hongxin Hu. Towards more practical threat models in artificial intelligence security. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.
- [9] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. Dos and don’ts of machine learning in computer security. In *Proceedings of the 31st USENIX Security Symposium*, 2022.
- [10] Guanhong Tao et al. Hard-label black-box universal adversarial patch attack. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.
- [11] Yuying Wan et al. BounceAttack: A query-efficient decision-based black-box attack. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy*, 2024.
- [12] Klim Kireev et al. On the robustness of machine learning models beyond adversarial settings. In *Proceedings of the 2023 Network and Distributed System Security Symposium*, 2023.
- [13] Chong Xiang, Saeed Mahloujifar, and Prateek Mittal. PatchCleanser: Certifiably robust defense against adversarial patches for any image classifier. In *Proceedings of the 31st USENIX Security Symposium*, 2022.

- [14] Chong Xiang, Tong Wu, Sihui Dai, Jonathan Petit, Suman Jana, and Prateek Mittal. PATCHCURE: Improving certifiable robustness, model utility, and computation efficiency of adversarial patch defenses. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.
- [15] Shimaa Ahmed, Ilia Shumailov, Nicolas Papernot, and Kassem Fawaz. Towards more robust keyword spotting for voice assistants. In *Proceedings of the 31st USENIX Security Symposium*, 2022.
- [16] Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zheng Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. ML-Doctor: Holistic risk assessment of inference attacks against machine learning models. In *Proceedings of the 31st USENIX Security Symposium*, 2022.
- [17] Xinyu Tang, Saeed Mahloujifar, Liwei Song, Virat Shejwalkar, Milad Nasr, Amir Houmansadr, and Prateek Mittal. Mitigating membership inference attacks by self-distillation through a novel ensemble architecture. In *Proceedings of the 31st USENIX Security Symposium*, 2022.
- [18] Jiacheng Li et al. MIST: Defending against membership inference attacks through membership-invariant subspace training. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.
- [19] Dingwen Wang et al. CAMP in the odyssey: Provably robust reinforcement learning with certified radius maximization. In *Proceedings of the 34th USENIX Security Symposium*, 2025.
- [20] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *Proceedings of the 2022 IEEE Symposium on Security and Privacy*, 2022.
- [21] Yueqi Ma et al. Avara: Measuring human perception of adversarial traffic sign examples using virtual and augmented reality. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, 2024.
- [22] Yaomin Jia et al. Physical adversarial attack on vehicle detector in the carla simulation environment. In *Proceedings of the 2022 Network and Distributed System Security Symposium*, 2022.
- [23] Siyuan Liu et al. X-Adv: Physical adversarial object attacks against x-ray prohibited item detection. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.
- [24] Yulong Cao et al. You can't see me: Physical removal attacks on lidar-based autonomous vehicles driving frameworks. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.
- [25] Shimaa Ahmed et al. Tubes among us: Analog attack on automatic speaker identification. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.
- [26] Ryunosuke Kobayashi et al. Invisible but detected: Physical adversarial shadow attack and defense on lidar object detection. In *Proceedings of the 34th USENIX Security Symposium*, 2025.
- [27] Yizheng Zhang et al. ATKSCOPES: Multiresolution adversarial perturbation as a unified attack on perceptual hashing and beyond. In *Proceedings of the 34th USENIX Security Symposium*, 2025.
- [28] Edoardo Debenedetti et al. Privacy side channels in machine learning systems. In *Proceedings of the 33rd USENIX Security Symposium*, 2024.

- [29] Milad Nasr et al. Exploiting component vulnerabilities in production ML pipelines. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, 2025.
- [30] Yanzuo Chen et al. OBSAN: An out-of-bound sanitizer to harden DNN executables. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.
- [31] R Spencer Hallyburton, Yupei Liu, Yulong Cao, Z Morley Mao, and Miroslav Pajic. Security analysis of camera-lidar fusion against black-box attacks on autonomous vehicles. In *Proceedings of the 31st USENIX Security Symposium*, 2022.
- [32] Lu Wang et al. From purity to peril: Backdooring merged models from “harmless” benign components. In *Proceedings of the 34th USENIX Security Symposium*, 2025.
- [33] Zhen Guo et al. Persistent backdoor attacks in continual learning. In *Proceedings of the 34th USENIX Security Symposium*, 2025.
- [34] Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y Zhao. Black-light: Scalable defense for neural networks against query-based black-box attacks. In *Proceedings of the 31st USENIX Security Symposium*, 2022.
- [35] Shenyi Zhang et al. JBShield: Defending large language models from jailbreak attacks through activated concept analysis and manipulation. In *Proceedings of the 34th USENIX Security Symposium*, 2025.
- [36] Zhiyuan Zhang et al. SafeSpeech: Robust and universal voice protection against malicious speech synthesis. In *Proceedings of the 34th USENIX Security Symposium*, 2025.
- [37] Dusan Popovic et al. DeBackdoor: A deductive framework for detecting backdoor attacks on deep models with limited data. In *Proceedings of the 34th USENIX Security Symposium*, 2025.
- [38] Tianshuo Cong et al. SSLGuard: A watermarking scheme for self-supervised learning pre-trained encoders. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022.
- [39] Yifan Lu et al. Neural dehydration: Effective erasure of black-box watermarks from DNNs with limited data. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, 2024.
- [40] Jaron Mink et al. “security is not my field, i’m a stats guy”: A qualitative root cause analysis of barriers to adversarial machine learning defenses in industry. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.
- [41] Xinyue Shen et al. Dynamic attention-based approaches for llm robustness. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, 2024.
- [42] Shahbaz Rezaei et al. On the accuracy-privacy trade-off of deep ensembles. In *Proceedings of the 2023 IEEE Symposium on Security and Privacy*, 2023.
- [43] Linyi Li et al. SoK: Certified robustness for deep neural networks. In *Proceedings of the 2023 IEEE Symposium on Security and Privacy*, 2023.

- [44] Phillip Rieger et al. DeepSight: Mitigating backdoor attacks in federated learning through deep model inspection. In *Proceedings of the 2022 Network and Distributed System Security Symposium*, 2022.

A Complete Paper Analysis Dataset

The complete analysis of all 454 papers is available in the supplementary CSV file: `all_conferences_analysis_1.csv`.

The dataset includes the following columns for each paper:

- Year, Conference, Filename, Title, Authors
- G1 (Focus), G2 (Attack Type), G3 (ML Type), G4 (Data Domain)
- G5 (Economics), G6 (Code Release), G7 (Real System Testing)
- T1 (Threat Model), T2 (Training Data Access)
- Q1 (Gradient Requirements), Q2 (Query Budget), Q3 (Computation)
- Gap indicator flags and Traditional Score

Benchmark meanings (aligned with the CSV fields used in the table):

- G1 Focus: atk (attack), def (defense), both.
- G2 Attack Type: Evasion, Poisoning, Privacy, Multiple.
- G3 ML Type: DL, Traditional, Both.
- G4 Data Domain: Images, Text, Audio, Malware, Other.
- G5 Economics mentioned: YES/NO.
- G6 Code released: YES/NO.
- G7 Real system testing: YES/NO.
- T1 Threat model: White-box, Gray-box, Black-box.
- T2 Training data access: Full, Partial, None.
- Q1 Requires gradients: YES/NO.
- Q2 Query budget: High (>1000), Low (<1000), None.
- Q3 Computation: High (GPU), Low (CPU).
- Traditional_Score (Gap Score 0 - 6): sum of six impractical-assumption flags (higher = less practical).

For full per-paper details (all 24 columns), please see the CSV file. Below is an ultra-compact, bordered summary table focused on the most critical taxonomy fields. Binary fields are rendered as checkmarks (yes) or dashes (no); threat/access levels are abbreviated (W/B/G/P, F/P, H/L).

Table 1: Compact summary of papers (key taxonomy fields).

Year	Venue	Paper (1st author)	G1	G2	G3	G4	G5	G6	G7	T1	T2	Grad	Qry	Comp	WB	Trad.
2022	ACM	Using	def	Poisoning	DL	Images	✓	✓	-	G	F	✓	H	L	0	3
2022	ACM	Reverse-Engineering	atk	Evasion	DL	Audio	✓	✓	-	B	-	H	L	L	0	2
2022	ACM	Tianshuo	def	Privacy	DL	Images	✓	✓	-	B	-	L	L	L	0	1
2022	ACM	CISPA	atk	Privacy	DL	Other	✓	-	-	B	-	L	L	L	0	2
2022	ACM	CISPA	atk	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	ACM	School	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	ACM	Physical	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	ACM	Framework	def	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	ACM	Florian	atk	Poisoning	DL	Images	✓	-	-	W	F	✓	H	L	1	5
2022	ACM	via	atk	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	ACM	ExamplesforLeak	def	Evasion	DL	Images	✓	✓	-	W	P	✓	H	L	1	4
2022	ACM	Harnessing	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	ACM	Feature	atk	Privacy	DL	Other	✓	-	✓	B	-	L	L	L	0	1
2022	ACM	Imperial	atk	Privacy	Both	Other	✓	✓	-	B	-	H	L	L	0	2
2022	ACM	Learning	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	ACM	CISPA	atk	Evasion	DL	Text	✓	✓	-	B	-	H	L	L	0	2
2022	ACM	Amrita	def	Poisoning	DL	Images	✓	-	-	B	-	H	L	L	0	4
2022	ACM	A	def	Evasion	Both	Text	✓	✓	-	B	-	L	L	L	0	1
2022	ACM	CISPA	atk	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	ACM	ETH	def	Privacy	Both	Other	✓	-	-	B	-	L	L	L	0	2
2022	ACM	Hanqing	atk	Evasion	DL	Audio	✓	✓	-	W	F	✓	H	L	1	4
2022	ACM	Stevens	atk	Privacy	DL	Other	✓	✓	-	W	F	✓	H	L	1	4
2022	ACM	Are	atk	Privacy	DL	Other	✓	✓	-	W	F	✓	H	L	1	4
2022	ACM	When	atk	Evasion	DL	Audio	✓	✓	✓	B	-	H	H	H	0	1
2022	ACM	Learning	atk	Privacy	DL	Images	✓	✓	-	B	-	H	L	L	0	2
2022	ACM	Using	atk	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	ACM	Neural	atk	Evasion	DL	Text	✓	✓	-	B	P	✓	H	H	0	3
2022	ACM	CISPA	atk	Privacy	DL	Images	✓	✓	-	B	P	-	L	L	0	1
2022	ACM	A	atk	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	ACM	LPGNet:	def	Privacy	DL	Other	✓	✓	-	B	-	H	L	L	0	3
2022	IEEE	Encoders	atk	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	IEEE	Illinois	atk	Evasion	DL	Images	✓	✓	-	B	-	L	L	L	0	1
2022	IEEE	Graph	atk	Privacy	DL	Other	✓	✓	-	B	F	✓	H	L	0	3
2022	IEEE	Imperceptible	atk	Evasion	DL	Text	✓	✓	✓	B	-	L	L	L	0	0
2022	IEEE	Security	def		DL	Other	✓	✓	-	B	-	L	L	L	0	1
2022	IEEE	Jialuo	def		DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	IEEE	Yun	atk	Evasion	DL	Other	✓	✓	-	B	-	H	L	L	0	2
2022	IEEE	Production	atk	Poisoning	DL	Images	✓	-	-	W	P	✓	H	L	1	5
2022	IEEE	Classification	def		Both	Malware	✓	✓	-	B	-	L	L	L	0	1
2022	IEEE	in	def	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	IEEE	Markus	def	Privacy	Both	Other	✓	✓	-	B	-	L	L	L	0	1
2022	IEEE	Eugene	atk	Poisoning	DL	Text	✓	✓	-	B	-	H	H	H	0	3
2022	IEEE	"Adversarial	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	IEEE	Attacks	atk	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	IEEE	Weight	atk	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	IEEE	SoK:	def		Both	Other	✓	-	-	B	-	L	L	L	0	2
2022	IEEE	Borja	atk	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	IEEE	Transformer	atk	Evasion	DL	Text	✓	✓	-	W	F	✓	H	L	1	4
2022	IEEE	Property	atk	Poisoning	DL	Images	✓	✓	-	B	F	-	H	L	0	2
2022	IEEE	Yuhao	atk	Evasion	DL	Images	✓	✓	-	B	-	H	L	L	0	2
2022	NDS	School	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	NDS	Federated	def	Poisoning	DL	Text	✓	-	-	W	-	L	L	L	1	3
2022	NDS	Property	atk	Privacy	DL	Images	✓	✓	-	B	F	✓	H	L	0	3
2022	NDS	RamBoAttack:	atk	Evasion	DL	Images	✓	✓	-	B	-	H	H	H	0	2
2022	NDS	Shengwei	atk	Privacy	DL	Images	✓	✓	✓	W	F	✓	H	L	1	3
2022	NDS	Robustness	both	Poisoning	DL	Images	✓	-	-	W	F	✓	H	L	1	5
2022	NDS	Ahmed	atk	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	USENIX	for	def	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	USENIX	Vocal	def	Evasion	DL	Audio	✓	✓	-	B	-	L	L	L	0	1
2022	USENIX	31st	atk	Privacy	Both	Other	✓	-	-	W	F	✓	H	L	1	5
2022	USENIX	August	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	USENIX	sponsored	atk	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	USENIX	August	atk	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4

Continued on next page

Year	Venue	Paper (1st author)	G1	G2	G3	G4	G5	G6	G7	T1	T2	Grad	Qry	Comp	WB	Trad.
2022	USENIX	sponsored	atk	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	USENIX	sponsored	atk	Poisoning	DL	Text	✓	✓	-	W	F	✓	H	L	1	4
2022	USENIX	Attacks	atk	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	USENIX	sponsored	def	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	USENIX	August	atk	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2022	USENIX	August	atk	Privacy	DL	Other	✓	✓	-	W	F	✓	H	L	1	4
2022	USENIX	Novel	atk	Privacy	Both	Other	✓	✓	-	B	—	-	H	L	0	2
2022	USENIX	Technische	def	—	Both	Malware	✓	-	-	B	—	-	L	0	2	
2022	USENIX	ICISPA	atk	Privacy	DL	Other	✓	✓	-	W	F	✓	H	L	1	4
2023	ACM	CISPA	def	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	ACM	Machine	def	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	ACM	Virginia	atk	Poisoning	DL	Images	✓	✓	-	G	P	✓	L	L	0	2
2023	ACM	through	atk	Privacy	DL	Audio	✓	✓	-	W	F	✓	H	L	1	4
2023	ACM	Stealing	atk	Evasion	DL	Text	✓	✓	-	B	—	-	H	L	0	2
2023	ACM	Stolen	both	Multiple	DL	Other	✓	✓	-	B	P	✓	H	L	0	3
2023	ACM	Information	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	ACM	Deep	def	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	ACM	Unforgeability	def	—	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	ACM	Stefano	def	Evasion	Traditional	Images	✓	✓	-	B	—	-	L	L	0	1
2023	ACM	Learning	atk	Poisoning	DL	Other	✓	✓	-	G	P	✓	H	L	0	3
2023	ACM	Secure	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	ACM	Detection	atk	Evasion	DL	Malware	✓	✓	-	B	—	-	L	L	0	1
2023	ACM	Code	def	—	DL	Other	✓	✓	-	G	—	-	L	L	0	1
2023	ACM	and	def	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	ACM	through	atk	Privacy	DL	Other	✓	✓	-	B	—	-	H	L	0	3
2023	ACM	Jingxuan	def	—	DL	Other	✓	✓	-	B	—	-	L	L	0	1
2023	ACM	and	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	ACM	Evading	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	ACM	Changqing	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	ACM	AntiFake:	def	Evasion	DL	Audio	✓	-	✓	B	—	-	H	L	0	2
2023	ACM	against	def	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	Effects	def	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	Deep	def	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	arXiv:2112.04558v2	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	Federated	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	SNAP:	atk	Poisoning	DL	Other	✓	✓	-	B	—	-	H	L	0	2
2023	IEEE	Federated	def	Poisoning	DL	Images	✓	-	-	W	F	✓	H	L	1	5
2023	IEEE	Information	atk	Privacy	DL	Text	✓	✓	-	B	—	-	H	L	0	2
2023	IEEE	Cover	atk	Poisoning	DL	Images	✓	✓	-	B	—	-	H	L	0	3
2023	IEEE	DepthFake	atk	Evasion	DL	Images	✓	✓	-	B	—	-	H	L	0	2
2023	IEEE	Jiameng	atk	Evasion	DL	Text	✓	✓	-	B	—	-	L	0	1	
2023	IEEE	Multi-party	def	Privacy	Both	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	Andrew	def	Privacy	Both	Other	✓	-	-	W	F	✓	H	L	1	5
2023	IEEE	Information	atk	Privacy	DL	Text	✓	✓	-	B	—	-	H	L	0	2
2023	IEEE	Extraction	atk	Evasion	DL	Images	✓	✓	-	B	—	-	H	L	0	2
2023	IEEE	Limin	atk	Poisoning	DL	Malware	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	Learning	def	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	Attacks	both	Evasion	DL	Text	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	Automatic	atk	Evasion	DL	Audio	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	yCybersecurity	def	—	DL	Images	✓	-	-	W	—	-	L	L	1	3
2023	IEEE	Ruijie	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	A	both	Privacy	DL	Images	✓	✓	-	B	P	-	H	L	0	2
2023	IEEE	SNAP:	atk	Poisoning	DL	Other	✓	-	-	B	—	-	H	L	0	3
2023	IEEE	Hong	def	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	BayBFed	def	Poisoning	DL	Images	✓	-	-	W	F	✓	H	L	1	5
2023	IEEE	Andre	atk	Evasion	DL	Audio	✓	✓	-	B	F	-	L	L	0	1
2023	IEEE	KASTEL	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	ELSA:	def	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	Shengwei	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	Patch-agnostic	def	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	RAB:	def	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	ETH	def	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	and	def	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	through	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	IEEE	StyleFool:	atk	Evasion	DL	Images	✓	✓	-	B	—	-	L	L	0	1
2023	NDS	DNN	def	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2023	NDS	Hadi	def	Evasion	DL	Audio	✓	✓	-	B	—	-	H	L	0	2

Continued on next page

Year	Venue	Paper (1st author)	G1	G2	G3	G4	G5	G6	G7	T1	T2	Grad	Qry	Comp	WB	Trad.	
2023	NDS	Yugeng	atk	Poisoning	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	NDS	Commercial	atk	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	NDS	Classifiers	def	Evasion	DL	Images ✓	✓	-	B	-	L	L	0	1			
2023	NDS	Learning	def	Evasion	DL	Other ✓	✓	-	B	-	H	L	0	3			
2023	NDS	Jiayun	atk	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	NDS	Klim	atk	Evasion	DL	Other ✓	-	-	B	-	H	H	0	3			
2023	NDS	Siyuan	atk	Poisoning	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	NDS	Universidad	def	Poisoning	DL	Text ✓	✓	-	W F	✓	H	L	1	4			
2023	NDS	Deep	atk	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	NDS	Chunyi	atk	Privacy	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	NDS	Robust	def	Poisoning	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	NDS	Physical	def	Evasion	Both	Other ✓	✓	✓	G P	-	L	L	0	0			
2023	NDS	*Technische	def	Privacy	Both	Text ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX		atk	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	Detection	atk	Evasion	DL	Malware✓	-	-	B	-	H	L	0	3			
2023	USENIX	Scientific	atk	Privacy	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	of	def		DL	Other ✓	✓	-	B	-	L	L	0	1			
2023	USENIX	Igor	atk	Poisoning	Both	Other ✓	-	-	W F	✓	H	L	1	5			
2023	USENIX	via	def	Privacy	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	FREE	def	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	TCISPA	def	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	Jonathan	atk	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	Vehicles	atk	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	Fairness	atk	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	Fairness	atk	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	Decompiling	atk	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	August	def		Both	Malware✓	✓	-	B	-	L	L	0	1			
2023	USENIX	Mazal	def	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	August	def		DL	Malware✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	is	atk	Privacy	DL	Images ✓	✓	-	B	P	✓	H	L	0	3		
2023	USENIX	is	def	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	Federated	def	Poisoning	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	Kunal	atk	Evasion	DL	Malware✓	✓	-	B	P	-	H	L	0	2		
2023	USENIX	Igor	def	Poisoning	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	by	def	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	is	atk	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	is	atk	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	August	atk	Poisoning	DL	Text ✓	-	-	B	-	H	L	0	3			
2023	USENIX	Universal	atk	Evasion	DL	Images ✓	✓	-	B	-	H	L	0	2			
2023	USENIX	is	atk	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2023	USENIX	August	def	Poisoning	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2024	ACM	from	atk	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2024	ACM	Xueluan	def	Privacy	DL	Images ✓	-	-	B	-	L	L	0	2			
2024	ACM	SafeEar:	def	Privacy	DL	Audio ✓	✓	-	B	-	L	L	0	1			
2024	ACM	Against	atk	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2024	ACM	Multi-Agent	atk	Evasion	DL	Images ✓	✓	-	B	-	H	H	0	2			
2024	ACM	Text-to-Image	def	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2024	ACM	Autonomous	def	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2024	ACM	Wei	atk	Poisoning	Both	Other ✓	✓	-	G P	✓	H	L	0	3			
2024	ACM	Byzantine-Robust	def	Poisoning	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2024	ACM	Recognition	atk	Evasion	DL	Audio ✓	✓	✓	B	-	H	0	0	0			
2024	ACM	Zijin	def	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2024	ACM	CSIRO's	atk	Poisoning	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2024	ACM	Items	atk	Poisoning	DL	Other ✓	✓	-	W F	✓	H	L	1	4			
2024	ACM	Applications	atk	Privacy	DL	Text ✓	✓	-	B	-	H	0	3				
2024	ACM	Attack	atk	Evasion	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2024	ACM	Attachment	atk	Evasion	DL	Malware✓	✓	✓	B	-	H	L	0	1			
2024	ACM	Information	atk	Evasion	DL	Text ✓	✓	-	W F	✓	H	L	1	4			
2024	ACM	Alchemy:	def	Evasion	DL	Images ✓	✓	-	W	-	H	L	1	4			
2024	ACM	S2NeRF:	both	Privacy	DL	Images ✓	✓	-	B	-	H	L	0	3			
2024	ACM	Certified	atk	Poisoning	DL	Other ✓	✓	-	W F	✓	H	L	1	4			
2024	ACM	ETH	atk	Privacy	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2024	ACM	Jie	atk	Privacy	DL	Images ✓	✓	-	B	P	-	H	L	0	2		
2024	ACM	Understanding	atk	Poisoning	DL	Images ✓	✓	-	W F	✓	H	L	1	4			
2024	ACM	jin	def	Evasion	DL	Malware✓	✓	-	W F	✓	H	L	1	4			
2024	ACM	A	def	Privacy	DL	Images ✓	✓	-	B	-	H	L	0	2			

Continued on next page

Year	Venue	Paper (1st author)	G1	G2	G3	G4	G5	G6	G7	T1	T2	Grad	Qry	Comp	WB	Trad.
2024	ACM	The Fisher into BadMerging: Optimization-based	def	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	ACM	Fisher into BadMerging: Optimization-based	def	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	ACM	BadMerging: Optimization-based	atk	Privacy	DL	Text	✓	✓	-	B	-	-	H	L	0	2
2024	ACM	BadMerging: Optimization-based	atk	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	ACM	Optimization-based	atk	Evasion	DL	Text	✓	✓	-	W	F	✓	H	L	1	4
2024	ACM	Optimization-based	atk	Evasion	DL	Text	✓	✓	-	W	F	✓	H	L	1	4
2024	ACM	Model Shenzhen	atk	Privacy	DL	Text	✓	✓	-	W	F	✓	H	L	1	4
2024	ACM	Information Learning	def	Evasion	DL	Images	✓	✓	-	B	-	-	H	L	0	1
2024	ACM	Blind Information	atk	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	ACM	Deepfake Membership	def	Privacy	DL	Text	✓	✓	-	B	-	-	L	L	0	2
2024	ACM	Membership	atk	Privacy	DL	Images	✓	✓	-	B	P	-	L	L	0	1
2024	ACM	Membership	atk	Privacy	DL	Images	✓	✓	-	B	P	-	L	L	0	1
2024	ACM	Neural	atk	Privacy	DL	Images	✓	✓	-	B	-	-	L	L	0	1
2024	ACM	of School Examples:	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	ACM	from Examples:	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	ACM	Institute Examples:	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	ACM	Institute Examples:	atk	Privacy	DL	Images	✓	✓	-	B	-	-	H	L	0	2
2024	ACM	of Examples:	atk	Poisoning	DL	Images	✓	✓	-	B	F	-	H	L	0	2
2024	ACM	School Examples:	atk	Evasion	DL	Text	✓	✓	-	W	F	✓	H	L	1	4
2024	ACM	Examples: Examples:	atk	Evasion	DL	Images	✓	✓	-	B	-	-	H	L	0	2
2024	ACM	Models Examples:	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	ACM	Learning Examples:	atk	Poisoning	DL	Images	✓	✓	-	B	P	-	L	L	0	1
2024	IEEE	Pattern	def	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	IEEE	Robin Yuzheng	def	Privacy	Both	Other	✓	✓	-	W	F	✓	H	L	1	4
2024	IEEE	Yuzheng	atk	Privacy	Both	Images	✓	✓	-	B	-	-	L	L	0	1
2024	IEEE	Joshua	atk	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	IEEE	Ziqi	def	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	IEEE	Linguistic	atk	Evasion	DL	Audio	✓	✓	✓	B	-	-	L	L	0	0
2024	IEEE	Language	atk	Evasion	DL	Text	✓	✓	-	W	F	✓	H	L	1	4
2024	IEEE	It's	def	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	IEEE	into	atk	Evasion	DL	Images	✓	✓	-	B	-	-	H	L	0	2
2024	IEEE	Test-Time	atk	Poisoning	DL	Images	✓	✓	-	G	P	✓	L	L	0	2
2024	IEEE	Xingshuo	atk	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	IEEE	Mohammad	atk	Poisoning	DL	Images	✓	-	-	G	P	✓	H	L	0	4
2024	IEEE	Attacks	def	Evasion	DL	Images	✓	✓	-	B	-	-	H	L	0	2
2024	IEEE	Distribution	atk	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	IEEE	Dropout	atk	Poisoning	DL	Images	✓	-	-	W	F	✓	H	L	1	5
2024	IEEE	GROV	def	Evasion	DL	Other	✓	✓	-	B	-	-	L	L	0	1
2024	IEEE	Mahmoud	atk	Evasion	DL	Other	✓	✓	-	B	P	-	L	L	0	1
2024	IEEE	Alex	def	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	IEEE	Accelerator	atk	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	IEEE	KASTEL	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	IEEE	Adversarial	def	Evasion	DL	Text	-	✓	-	W	-	-	L	L	1	3
2024	IEEE	2Services	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	NDS	Large	def	Evasion	DL	Text	✓	✓	-	W	F	✓	H	L	1	4
2024	NDS	DEMASQ	def	Evasion	DL	Text	✓	-	-	B	-	-	H	L	0	4
2024	NDS	Attacks	atk	Privacy	DL	Audio	✓	✓	-	W	F	✓	H	L	1	4
2024	NDS	Poisoning	atk	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	NDS	by	def	Privacy	DL	Images	✓	✓	-	B	-	-	L	L	0	1
2024	NDS	CamPro	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	NDS	Ensemble	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	NDS	Gelei	atk	Evasion	DL	Text	✓	-	-	B	-	-	H	H	0	3
2024	NDS	Watermarking	def	Evasion	DL	Audio	✓	✓	✓	W	F	-	L	L	1	1
2024	NDS	ICS	atk	Evasion	DL	Other	✓	✓	-	W	F	✓	H	L	1	4
2024	NDS	Federated	def	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	NDS	Chengkun	def	Poisoning	DL	Text	✓	✓	-	B	-	✓	L	L	0	2
2024	NDS	Recognition	atk	Evasion	DL	Audio	✓	✓	-	W	F	✓	H	L	1	4
2024	NDS	Transpose	atk	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	NDS	Inversion-based	def	Privacy	DL	Images	✓	✓	-	W	F	✓	H	L	1	4

Continued on next page

Year	Venue	Paper (1st author)	G1	G2	G3	G4	G5	G6	G7	T1	T2	Grad	Qry	Comp	WB	Trad.
2024	NDS	Networks	atk	Poisoning	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	NDS	Browser	def	Privacy	Traditional	Other ✓	-	-	B	-	-	L	L	0	2	
2024	NDS	Deep	def	Evasion	DL	Images ✓	✓	-	B	-	-	H	L	0	2	
2024	NDS	Bang	def	Privacy	DL	Other ✓	✓	-	B	-	-	L	L	0	1	
2024	NDS	Against	atk	Poisoning	DL	Images ✓	✓	-	B	P	✓	H	L	0	3	
2024	NDS	Speaker	atk	Evasion	DL	Audio ✓	✓	✓	B	-	-	L	L	0	0	
2024	NDS	Mitigating	def	Poisoning	DL	Images ✓	-	-	B	-	-	L	L	0	2	
2024	NDS	for	def		DL	Malware ✓	✓	-	B	-	-	L	L	0	1	
2024	NDS	in	both	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	NDS	Attacks	def	Poisoning	DL	Text ✓	✓	-	W	F	-	L	L	1	2	
2024	NDS	Chaoxiang	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Tong	atk	Evasion	DL	Text ✓	✓	-	B	-	-	L	L	0	1	
2024	USENIX	Attacks	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Approach	def	Poisoning	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Qingzhao	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Zhenghang	def		Both	Other ✓	✓	-	B	-	-	L	L	0	1	
2024	USENIX	Synthesis	def	Privacy	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Measuring	def		DL	Audio ✓	✓	-	B	-	-	L	L	0	1	
2024	USENIX	Large	atk	Evasion	DL	Text ✓	✓	-	B	-	-	H	L	0	2	
2024	USENIX	Efficiency	def	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Efficiency	def	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Kathrin	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	DNN-GP:	both	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	for	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	DeepEclipse	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	DeepEclipse	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Between	atk	Multiple	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Red-Teaming	atk	Evasion	DL	Images ✓	✓	-	B	-	-	H	L	0	2	
2024	USENIX	Red-Teaming	atk	Evasion	DL	Images ✓	✓	-	B	-	-	H	L	0	2	
2024	USENIX	Property	atk	Privacy	DL	Images ✓	✓	-	B	F	-	H	L	0	2	
2024	USENIX	Transposed	def		DL	Images ✓	-	-	W	F	✓	H	L	1	5	
2024	USENIX	in	def	Privacy	DL	Images ✓	✓	-	B	-	-	L	L	0	1	
2024	USENIX	Shuaifan	def	Privacy	DL	Images ✓	✓	✓	W	F	✓	H	L	1	3	
2024	USENIX	CISPA	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	CISPA	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Runtime	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Splitting	def	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Yixin	atk	Privacy	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Changjiang	atk	Poisoning	DL	Images ✓	✓	-	B	F	✓	H	L	0	3	
2024	USENIX	Guangsheng	atk	Privacy	DL	Images ✓	-	-	W	F	✓	H	L	1	5	
2024	USENIX	Bailey	def	Privacy	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Meng	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	"Mique	def	Privacy	DL	Images ✓	✓	-	W	l	✓	H	L	1	4	
2024	USENIX	CISPA	atk	Privacy	DL	Images ✓	✓	-	B	P	-	H	H	0	2	
2024	USENIX	False	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Shenchen	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	through	atk	Privacy	DL	Other ✓	-	-	B	-	✓	H	L	0	4	
2024	USENIX	Shaofeng	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Mudjacking:	def	Poisoning	DL	Images ✓	-	-	B	-	✓	H	L	0	4	
2024	USENIX	Randomised	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	33rd	atk	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	is	def		DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	August	def	Privacy	Both	Text ✓	✓	-	B	-	L	L	L	0	1	
2024	USENIX	August	atk	Privacy	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	is	atk	Poisoning	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	is	def		DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	is	def	Poisoning	DL	Images ✓	-	-	B	-	-	L	L	0	2	
2024	USENIX	Attacks	def	Privacy	DL	Images ✓	✓	-	B	-	-	L	L	0	3	
2024	USENIX	Model	atk	Evasion	DL	Text ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	Zooming	both	Evasion	DL	Images ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	August	atk	Privacy	DL	Text ✓	✓	-	B	-	-	L	L	0	1	
2024	USENIX	is	def		DL	Text ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	August	def		DL	Other ✓	✓	-	W	F	-	H	L	1	3	
2024	USENIX	August	def		DL	Other ✓	✓	-	W	F	✓	H	L	1	4	
2024	USENIX	August	def		Evasion	DL	Images ✓	✓	✓	B	-	-	L	L	0	1
2024	USENIX	August	def		Both	Malware ✓	-	-	B	-	-	L	L	0	2	
2024	USENIX	Learning	def	Evasion	Both	Other ✓	✓	✓	-	W	F	✓	H	L	1	4

Continued on next page

Year	Venue	Paper (1st author)	G1	G2	G3	G4	G5	G6	G7	T1	T2	Grad	Qry	Comp	WB	Trad.
2024	USENIX	Framework	def		DL	Text	✓	✓	-	B	-		L	0	1	
2024	USENIX		def		DL	Images	✓	✓	-	B	-		L	0	1	
2024	USENIX		atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2024	USENIX	Inference	def	Privacy	DL	Images	✓	-	-	G	-	✓	H	L	0	4
2025	ACM	ITechnion	atk	Evasion	DL	Text	✓	✓	-	B	-		H	L	0	2
2025	ACM	in	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2025	ACM	arXiv:2501.05928v2	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4	
2025	ACM	Anti-Facial	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2025	ACM	BIFOLD	atk	Evasion	DL	Other	✓	✓	-	W	F	✓	H	H	1	4
2025	ACM	Targeted	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2025	ACM	Language	atk	Poisoning	DL	Text	✓	✓	-	B	-	L	L	0	1	
2025	ACM	Models	def	Evasion	DL	Images	✓	-	✓	G	P	✓	H	H	0	3
2025	ACM	Adversarial	atk	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2025	ACM	Construction	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2025	ACM	Latent-based	atk	Evasion	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2025	ACM	Milad	atk	Evasion	DL	Malware	✓	✓	✓	W	F	✓	H	L	1	3
2025	ACM	VillainNet:	atk	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2025	ACM	Deep	def		DL	Malware	✓	✓	-	B	-		L	0	1	
2025	USENIX	Jiachen	def	Evasion	DL	Images	✓	✓	✓	B	-	✓	H	H	0	2
2025	USENIX	Lingchen	def	Evasion	DL	Text	✓	✓	-	W	F	-	L	L	1	2
2025	USENIX	Revisiting	atk	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2025	USENIX	Zhisheng	def	Privacy	DL	Audio	✓	✓	-	B	-	L	L	0	1	
2025	USENIX	with	def	Evasion	DL	Other	✓	✓	-	W	F	-	H	L	1	3
2025	USENIX	Persistent	atk	Poisoning	DL	Images	✓	✓	-	W	-	✓	H	L	1	4
2025	USENIX	Benign	atk	Poisoning	DL	Images	✓	✓	-	W	F	✓	H	L	1	4
2025	USENIX	34th	def	Evasion	DL	Malware	✓	✓	-	B	-	✓	H	L	0	3