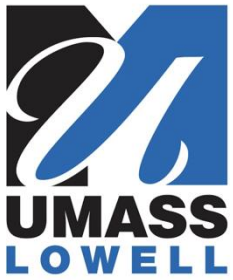
Network and Distributed System Security 2026

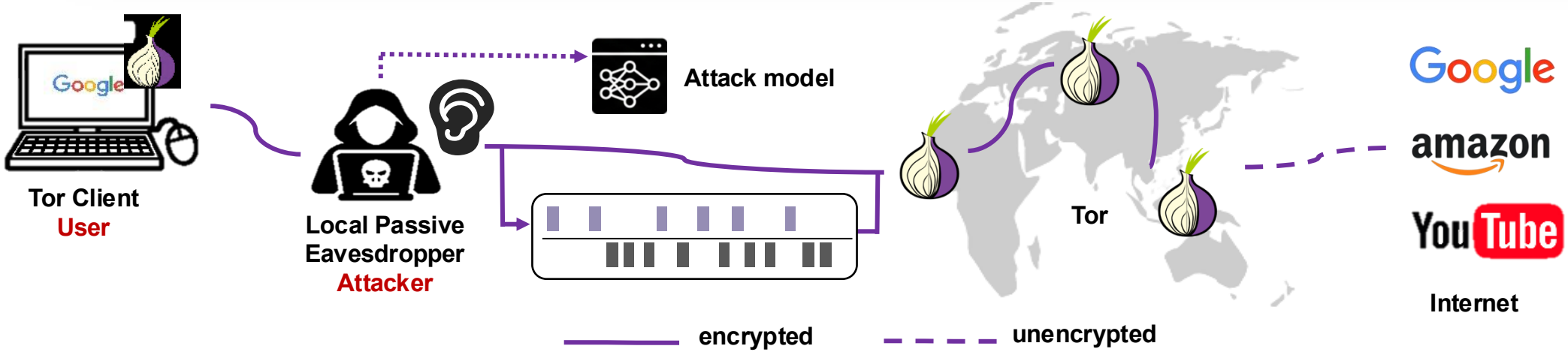
Cease at the Ultimate Goodness:

Towards Efficient Website Fingerprinting Defense via Iterative Mutual Information Minimization

Rong Wang†, Zhen Ling*†, Guangchi Liu†, Shaofeng Li†, Junzhou Luo†‡, Xinwen Fu§
† Southeast University
‡ Fuyao University of Science and Technology
§ University of Massachusetts Lowell

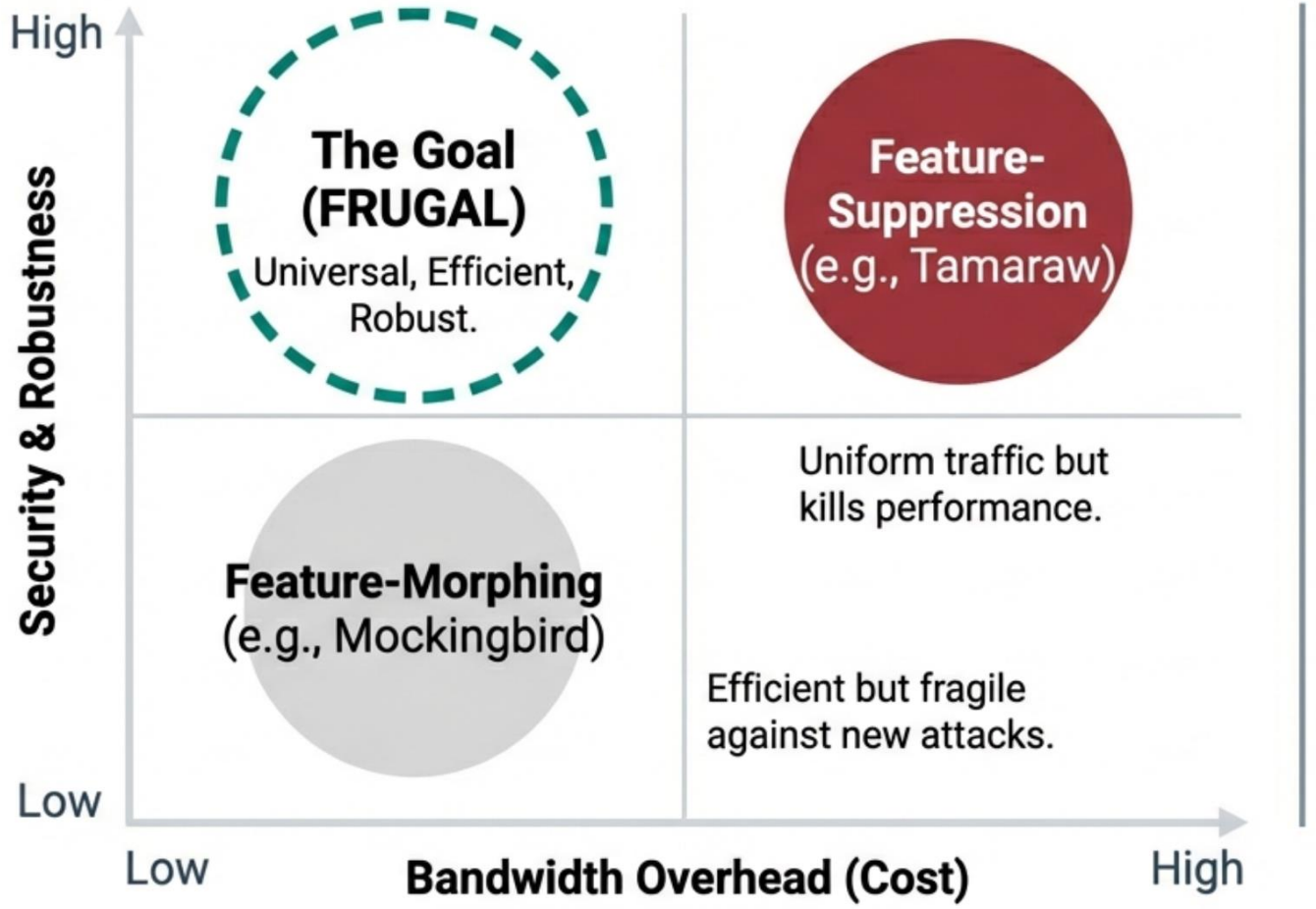


The Invisible Threat in Anonymity Networks



- **What is Tor?** Tor is the leading anonymity network that conceals a user's online activity by routing traffic through a worldwide volunteer relay system.
- **The Vulnerability:** Even with anonymity protection, traffic patterns (packet size, direction, and timing) remain visible to local eavesdroppers.
- **The Consequence:** Deep Learning attackers can exploit these patterns as "fingerprints", achieving over 98% accuracy in de-anonymizing user visits.

The Defense Dilemma



Trilemma of Current Defenses:

- Attack Model Agnostic: Morphing fails against unknown attackers.
- Bandwidth Efficiency: Suppression causes congestion and latency.
- Adversarial Robustness: Defenses leak information when attackers retrain.

Core Philosophy

From Deception to Elimination

The Problem with Previous Methods:

Most existing traffic morphing-based defenses focus on deceiving specific classifiers.

A Data-Centric Approach:

FRUGAL shifts focus from deceiving models to fundamentally eliminating the identifiable patterns within the traffic itself.

Trace A $[+1, +1, -1]$

Trace B $[+1, -1, +1]$

LCS/padding →

Indistinguishable Version

Trace A' $[+1, +1, -1, +1]$

Trace B' $[+1, +1, -1, +1]$

High Security, but Excessive Overhead !

The FRUGAL Insight: Instead of total elimination, we need a way to measure and minimize how much information a sample x leaks about its label .

The Paradigm Shift: Mutual Information Minimization

From Deception to Elimination

Metric:

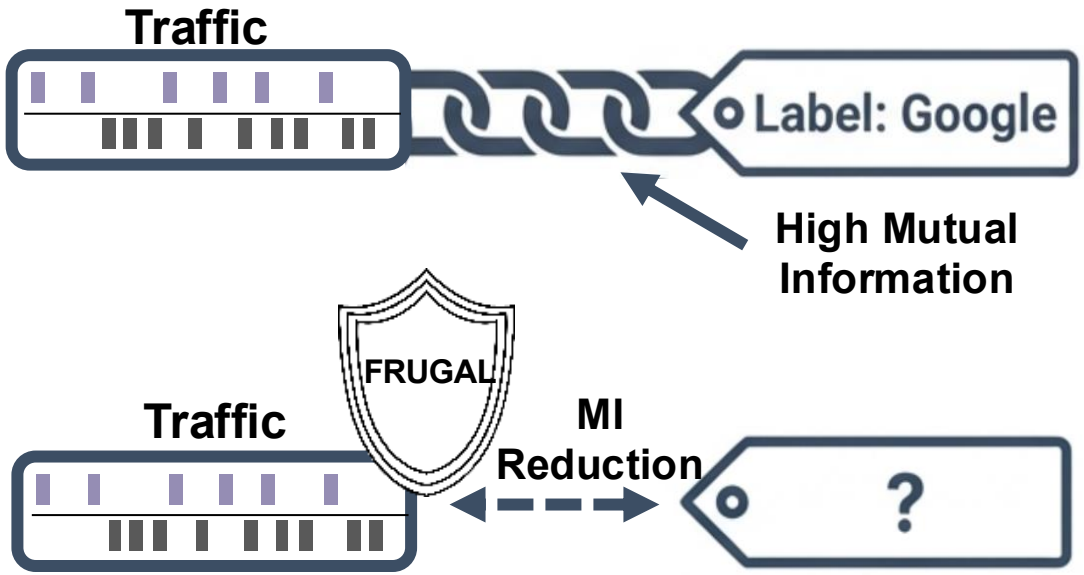
Mutual Information(MI)

$$I(x\,;\,y) = H(y) - H(y|x)$$

Goal:

Minimize the shared information between Traffic(x) and Website Labels (y).

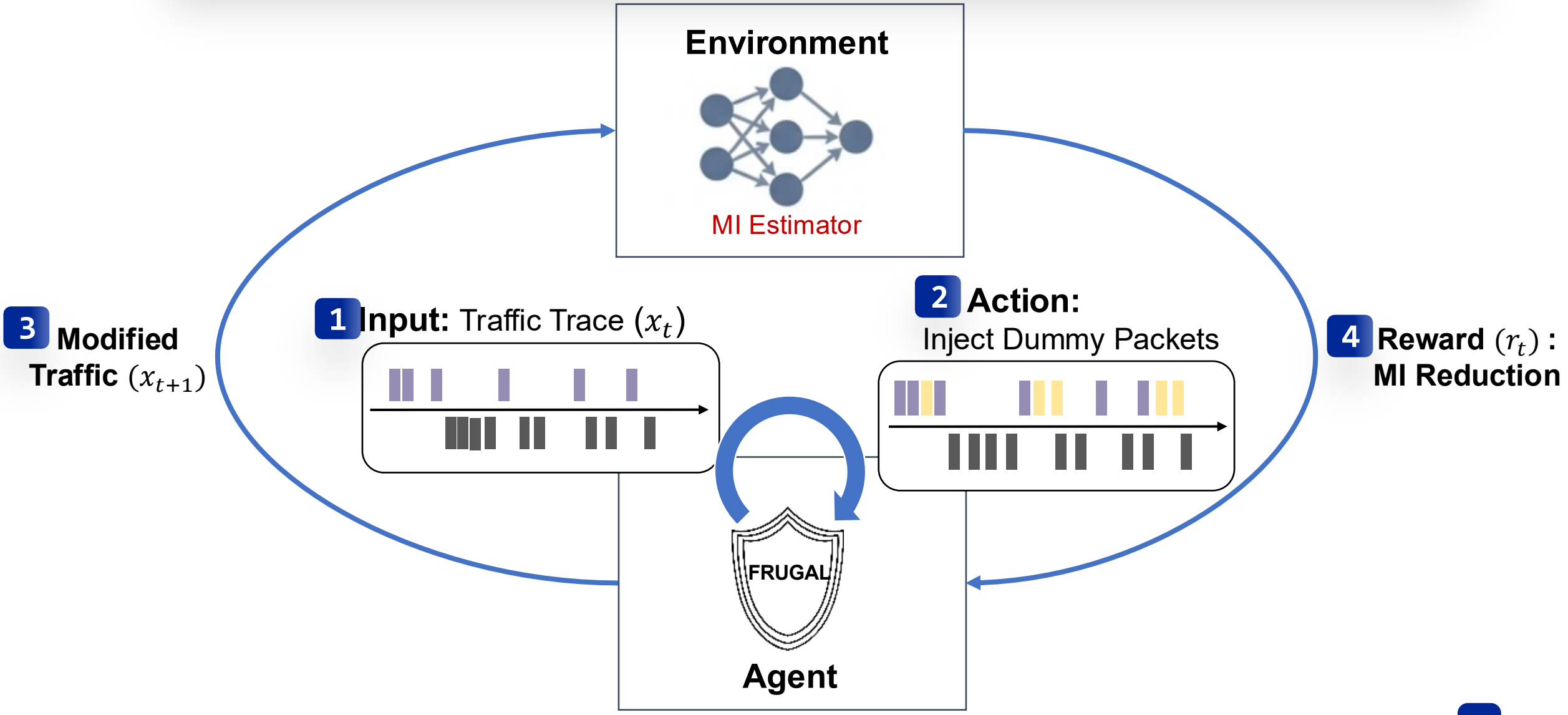
$$\min I(x;y) \leftrightarrow max\ H(\mathrm{y}|x)$$



Mechanism :

Strategic injection of dummy packets to maximize attacker uncertainty (Entropy).

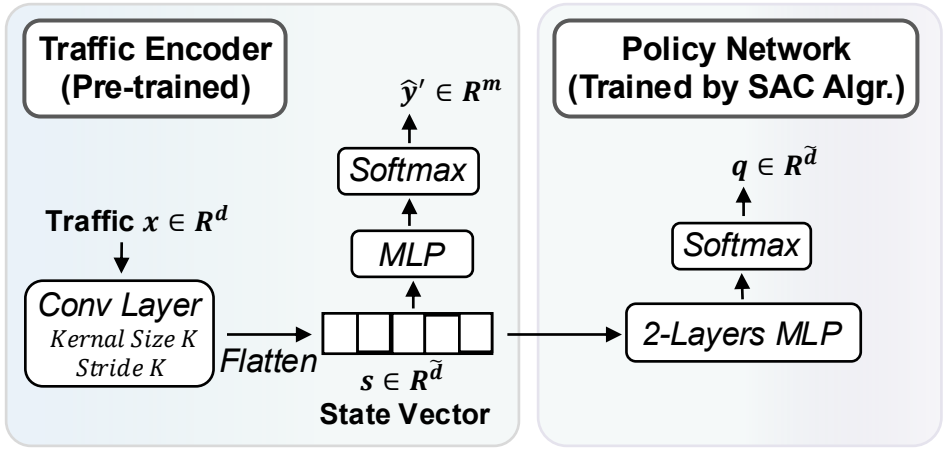
The Architecture: Reinforcement Learning Loop



The Agent: Smart Injection Policy

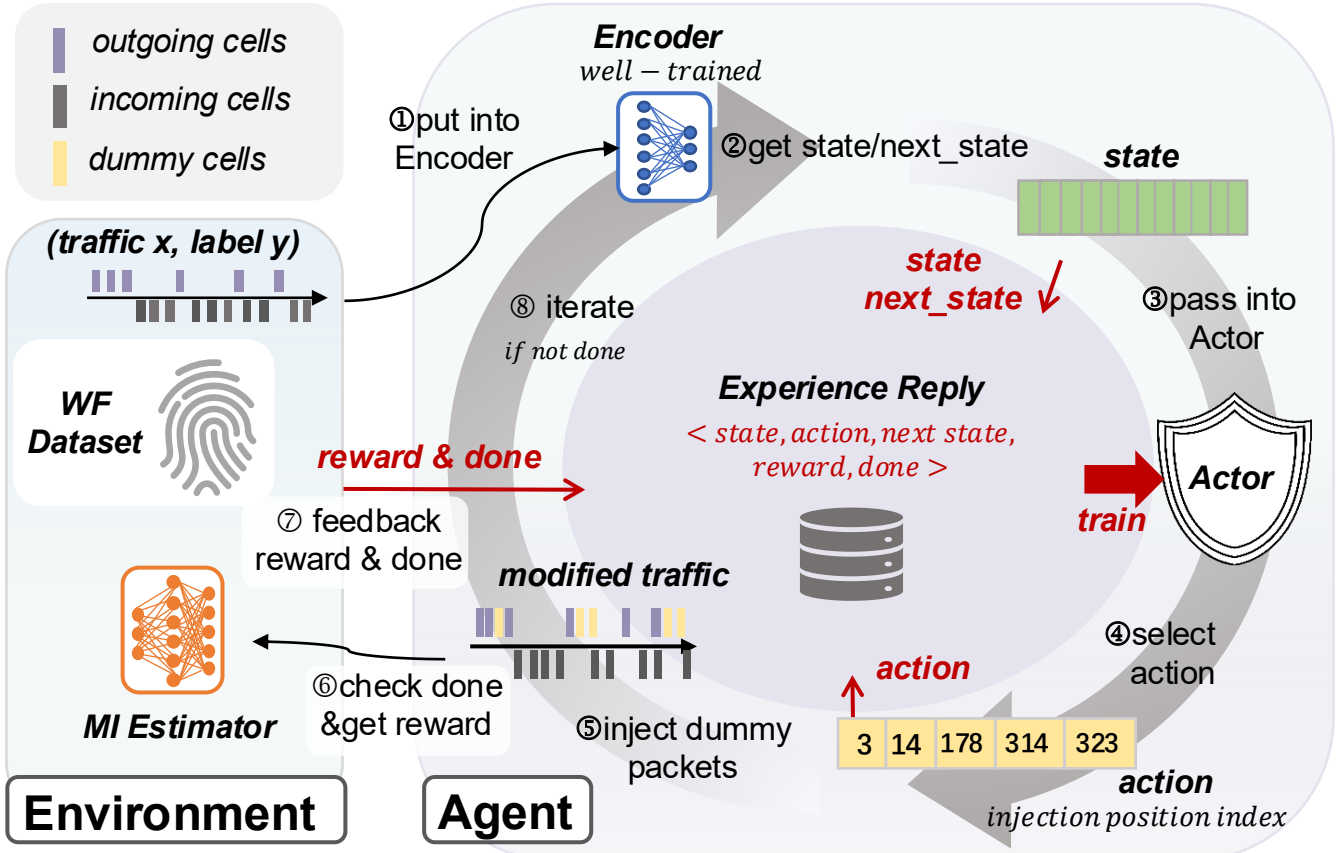
Overcoming the Curse of Dimensionality

The Architecture of Agent



Traffic Encoder shrinks the search space to $\frac{1}{K}$ of the original trace length while preserving key classification features.

Training Process of FRUGAL



The Environment: The CLUB Estimator

Rewarding Uncertainty

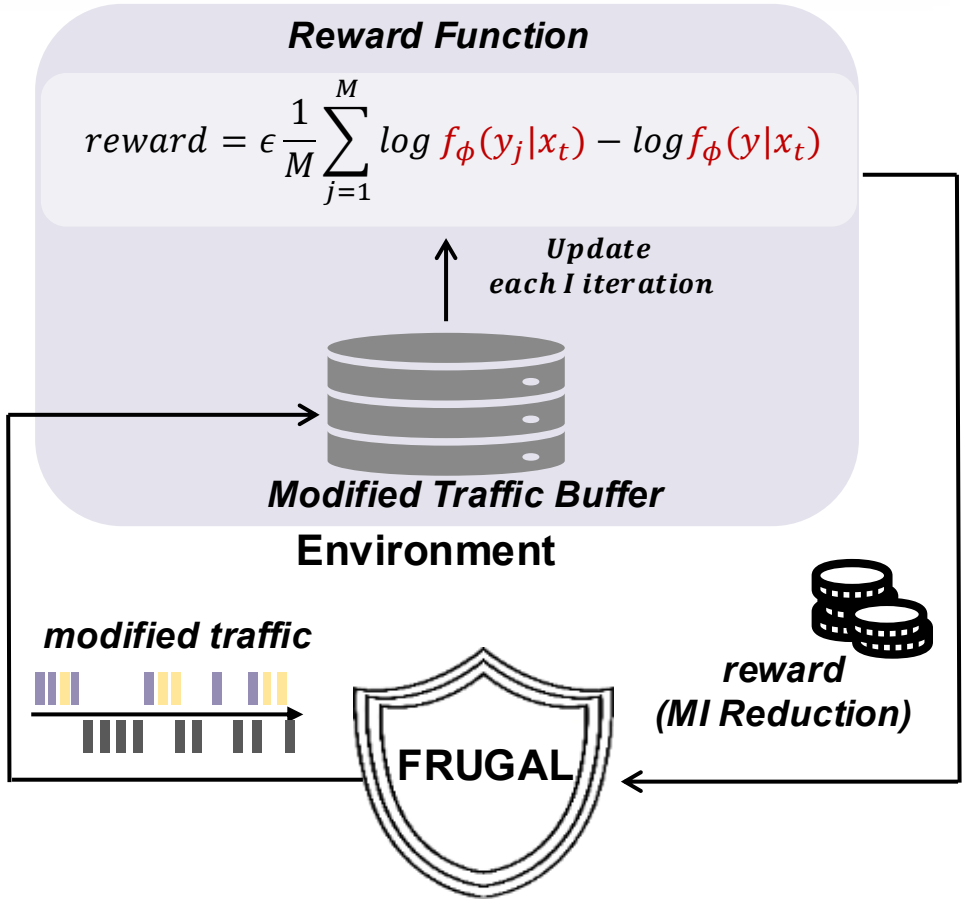
Objective: Guide the agent to minimize Mutual Information.

Tools: CLUB (Contrastive Log-ratio Upper Bound) Estimator.

Reward Function:

$$\mathrm{R}(x_t) = -\log \mathrm{f}_\phi(\mathrm{y}|\mathrm{x}_\mathrm{t}) + \epsilon \frac{1}{\mathrm{M}} \sum_{\mathrm{j}=1}^{\mathrm{M}} \log \mathrm{f}_\phi(\mathrm{y}_\mathrm{j}|\mathrm{x}_\mathrm{t})$$

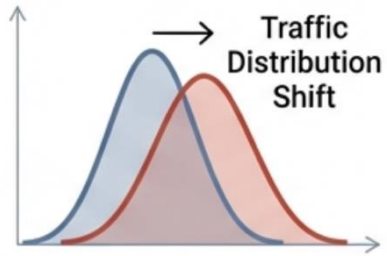
Interpretation: Minimize confidence in the correct site, maximize confusion with other sites.



Robustness: Dynamic Feature Elimination (DFE)

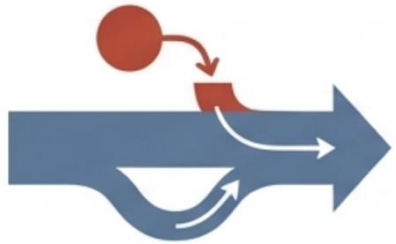
Defeating Adversarial Training via Iterative Updates

Solving "Estimator Drift"



Packet injections cause a distribution shift, rendering static MI estimators inaccurate over time.

Adaptive Judging

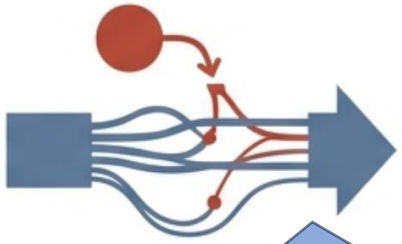


DFE periodically retrains the classifier f_ϕ on modified traffic every I iterations.

CMI Optimization

This updates the enviro-nment into a CMI estimator, targeting the most informative residual patterns.

Theoretical Guarantee

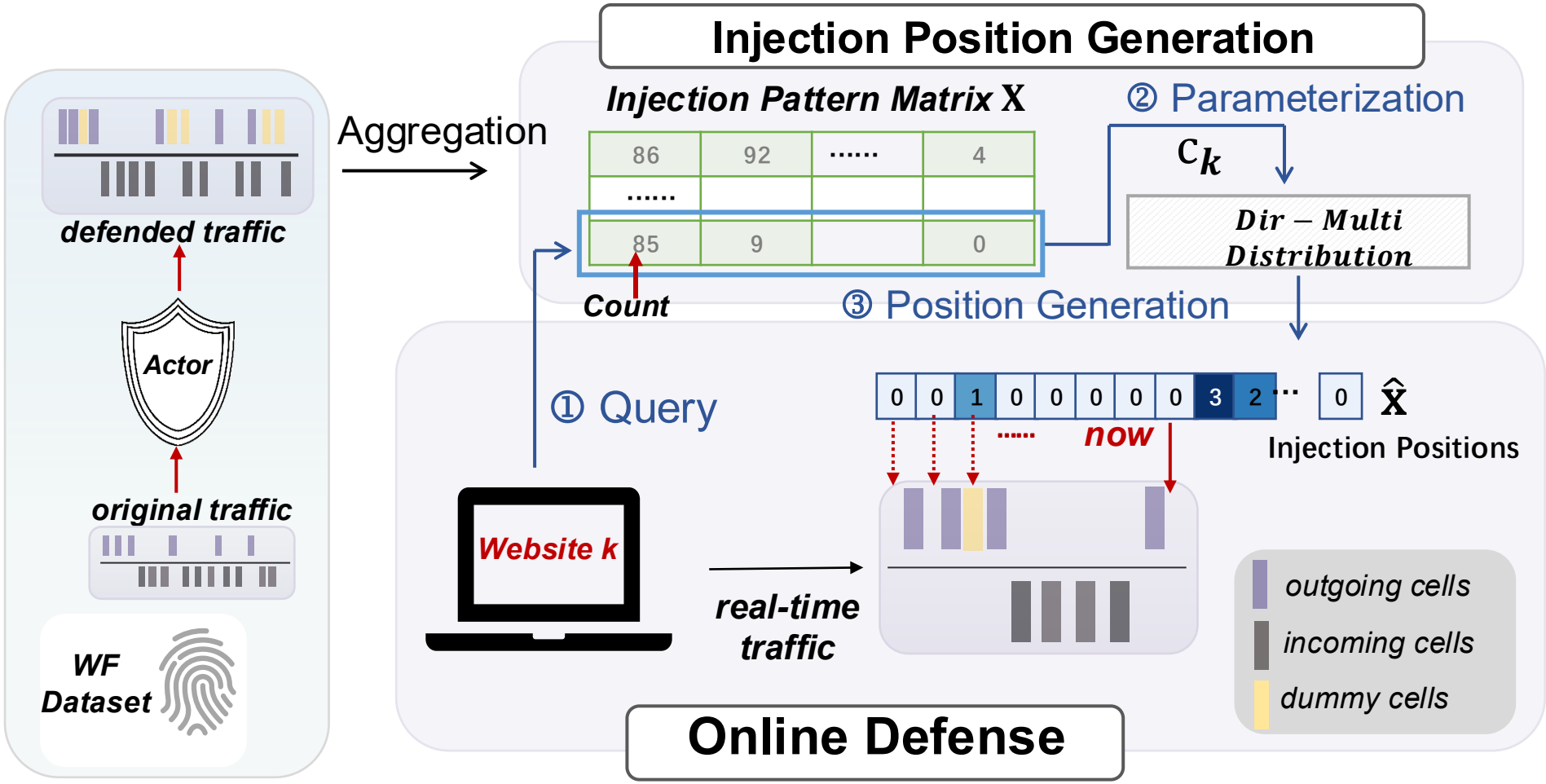


A greedy position selection strategy guided by a dynamic CMI estimator is proven to achieve the global maximum MI reduction.

By constantly updating the CMI Estimator, the Agent is forced to hide even the deepest, most subtle patterns, effectively immunizing itself against future adversarial training.

FRUGAL-Online: Real-Time Deployment

Distilling Offline Intelligence for Online Speed

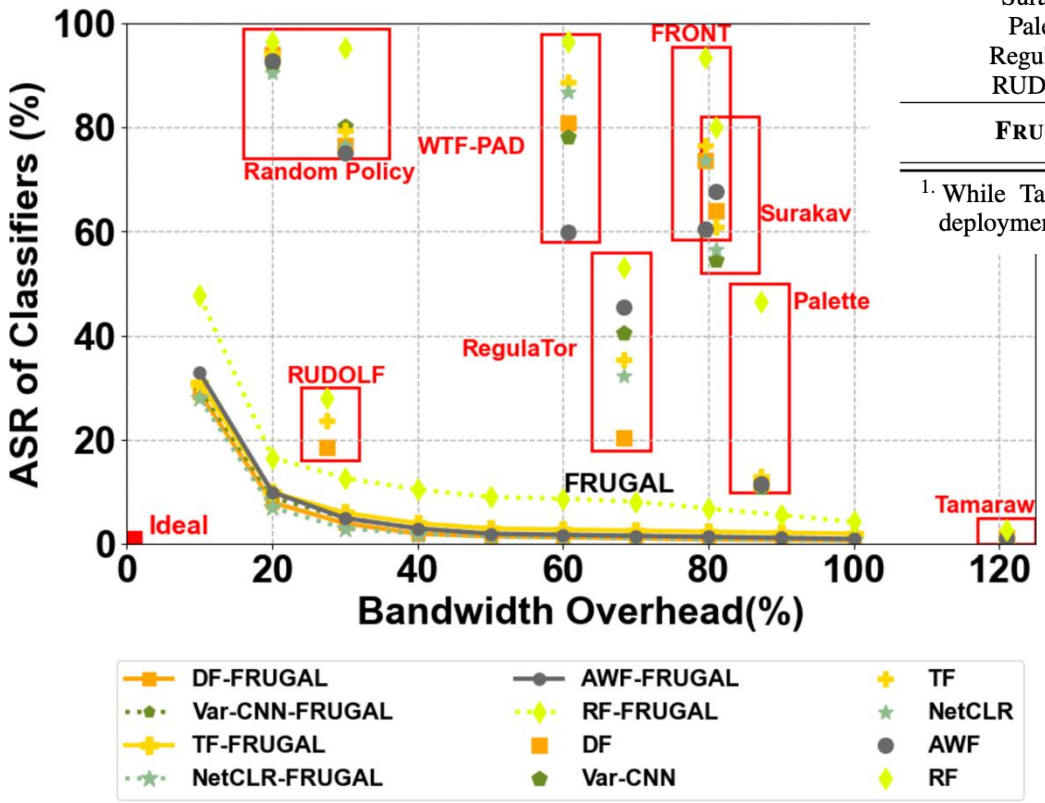


Evaluation Setup

Dataset (DF Dataset)	Attack Models (The Adversaries)	Competitors (The Baselines)
• 95 * 1000 Monitored Websites(Alexa Top) • 40,000 Unmonitored Websites • Realistic, dynamic network conditions	DF Var-CNN NetCLR TF AWF RF	Palette FRONT WTF-PAD Tamaraw RUDOLF

Metrics: Attack Success Rate (ASR) vs. Bandwidth Overhead (BWO).

Result: Closed World



PERFORMANCE IN THE CLOSED-WORLD SCENARIO

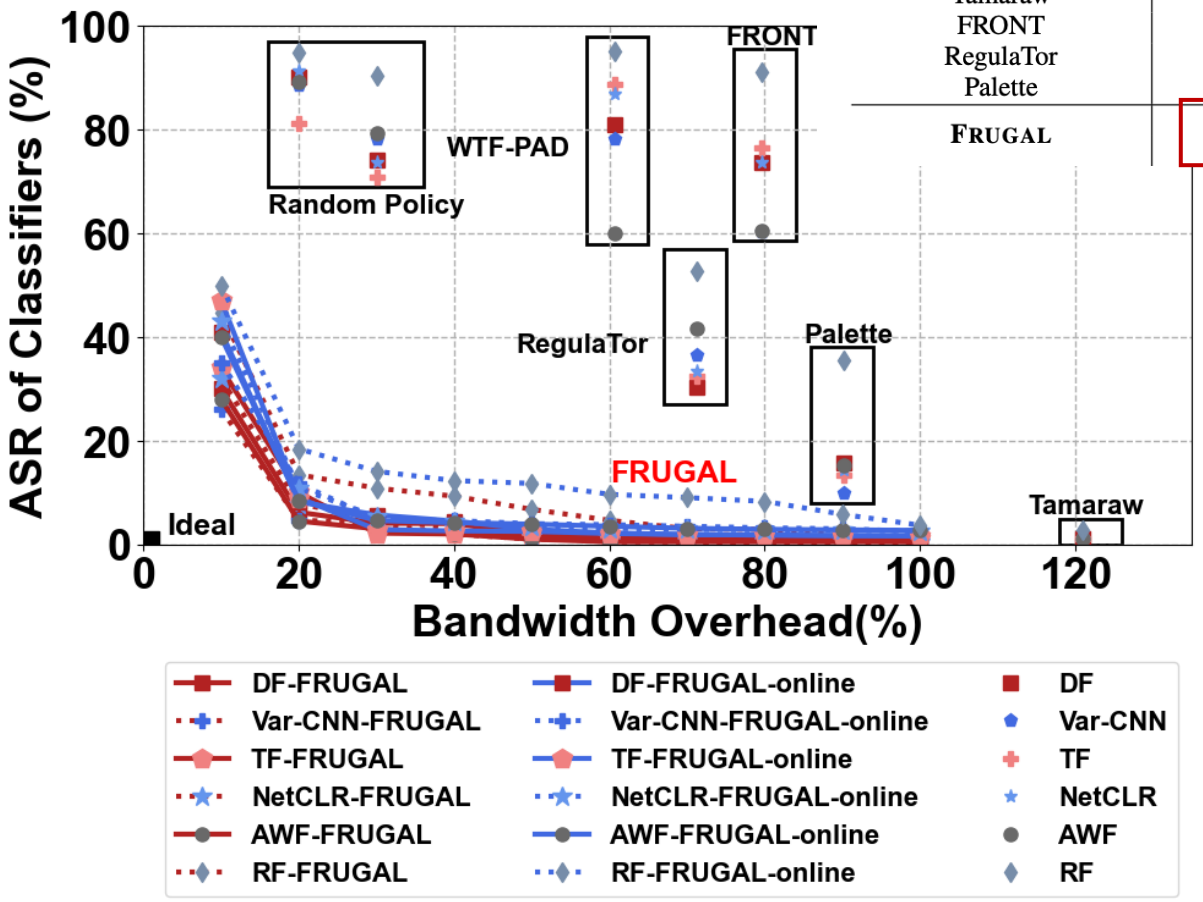
Defenses	BWO	ASR					
		DF	Var-CNN	NetCLR	TF	AWF	RF
Random Injection	20%	93.98%	91.98%	90.6%	94.3%	92.76%	96.58
	30%	76.59%	80.17%	76.59%	79.34%	75.19%	95.3
WTF-PAD	60.7%	80.92%	78.14%	86.92%	88.65%	59.96%	96.58%
Tamaraw[1]	121%	1.05%	0.98%	1.01%	1.12%	1.05%	2.09%
FRONT	79.6%	73.62%	60.25%	73.62%	76.46%	60.44%	93.34%
Surakav	81%	64%	54.6%	56.69%	60.95%	67.65%	79.94%
Palette	87.17%	11.54%	10.99%	11.2%	12.91%	11.54%	46.43%
RegulaTor	68.3%	20.41%	40.52%	32.31%	35.52%	45.6%	53.11%
RUDOLF	27.46%	18.59%	-	-	23.71%	-	28%
FRUGAL	**20%**	**6.87%**	**8.03%**	**12.73%**	**10.37%**	**10.12%**	**16.6%**
	30%	**2.68%**	**2.61%**	**6.68%**	**5.67%**	**5.73%**	**12.7%**

1. While Tamaraw achieves commendable performance on benchmark datasets, its excessive bandwidth overhead renders it impractical for deployment in anonymous network environments where performance and user experience are critical.

- FRUGAL:<3% ASR with only 30% BWO
- Palette (SOTA): ~11.5% ASR with 87% BWO

FRUGAL consistently occupies the "ideal" bottom-left corner of the ASR-BWO trade-off curve across all bandwidth levels.

Result: Open World & One-page



PERFORMANCE IN THE OPEN-WORLD SCENARIO

Defenses	BWO	ASR					
		DF	Var-CNN	NetCLR	TF	AWF	RF
Random Injection	20%	90.12%	88.6%	91.43%	81.3%	89.3%	94.8%
	30%	74.04%	78.25%	70.63%	73.75%	79.4%	90.3%
WTF-PAD	60.7%	80.92%	78.14%	86.92%	88.65%	59.96%	95.12%
Tamaraw	121%	1.02%	0.9%	1.0%	1.12%	0.95%	2.07%
FRONT	99%	57.23%	50.25%	54.62%	56.46%	57.44%	91.2%
RegulaTor	71.32%	30.41%	36.52%	33.5%	32.12%	41.6%	52.61%
Palette	90.2%	15.81%	9.89%	14.31%	13.41%	15.32%	35.42%
FRUGAL	**20%**	**6.2%**	**6.55%**	**7.8%**	**5.7%**	**4.5%**	**13.43%**
	30%	**4.09%**	**4.7%**	**3%**	**2.17%**	**2.58%**	**10.85%**

FRUGAL performs well even in a tougher evaluation setting.

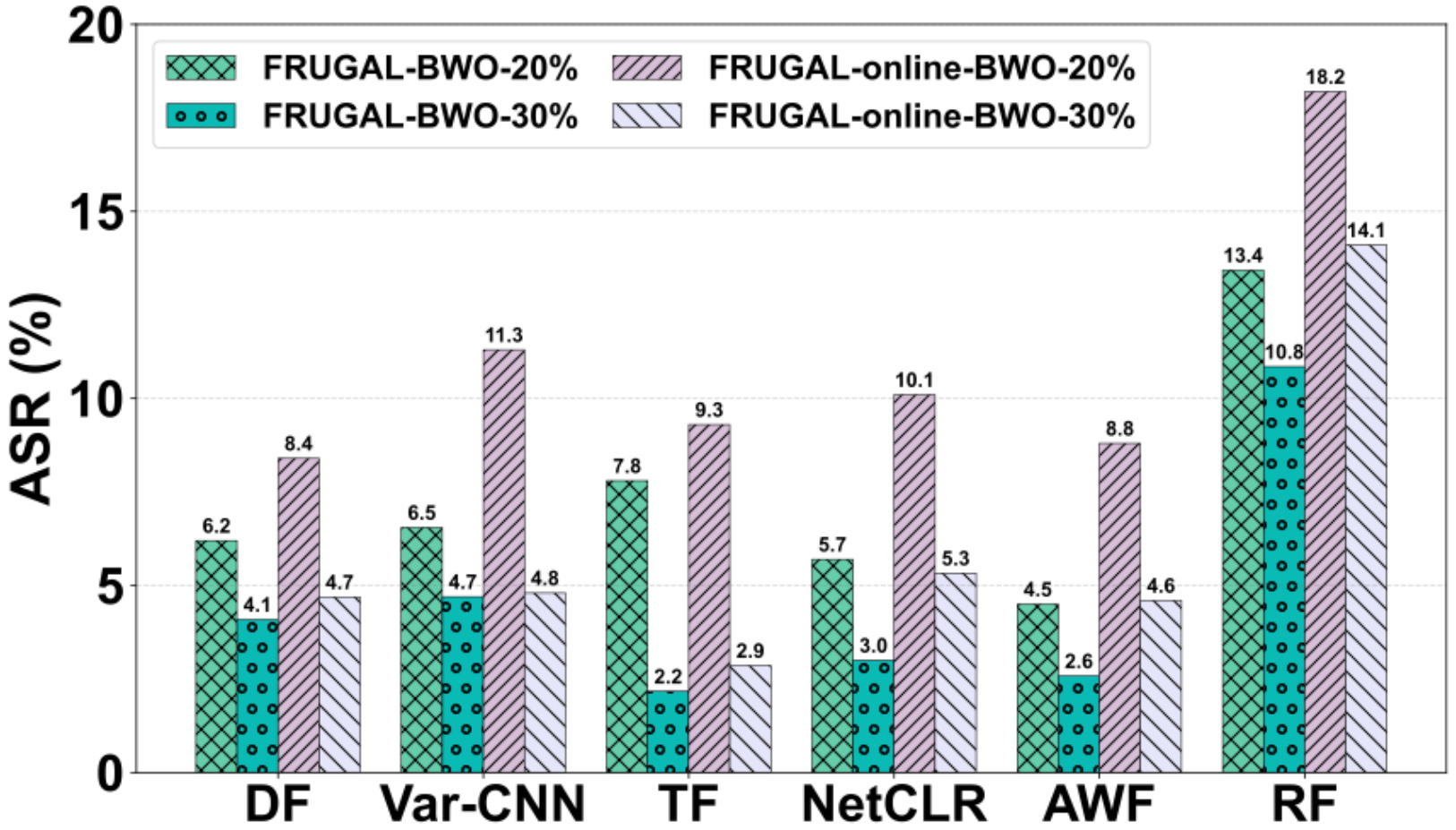
EVALUATION OF FRUGAL IN THE ONE-PAGE SETTING COMPARED TO OTHER DEFENSE METHODS

	Defenses			
	FRUGAL	Palette	RUDOLF	RegulaTor
BWO	**19.63%**	109.17%	27.46%	48.3%
Average ASR	**6.54%**	36.85%	67.3%	55.71%

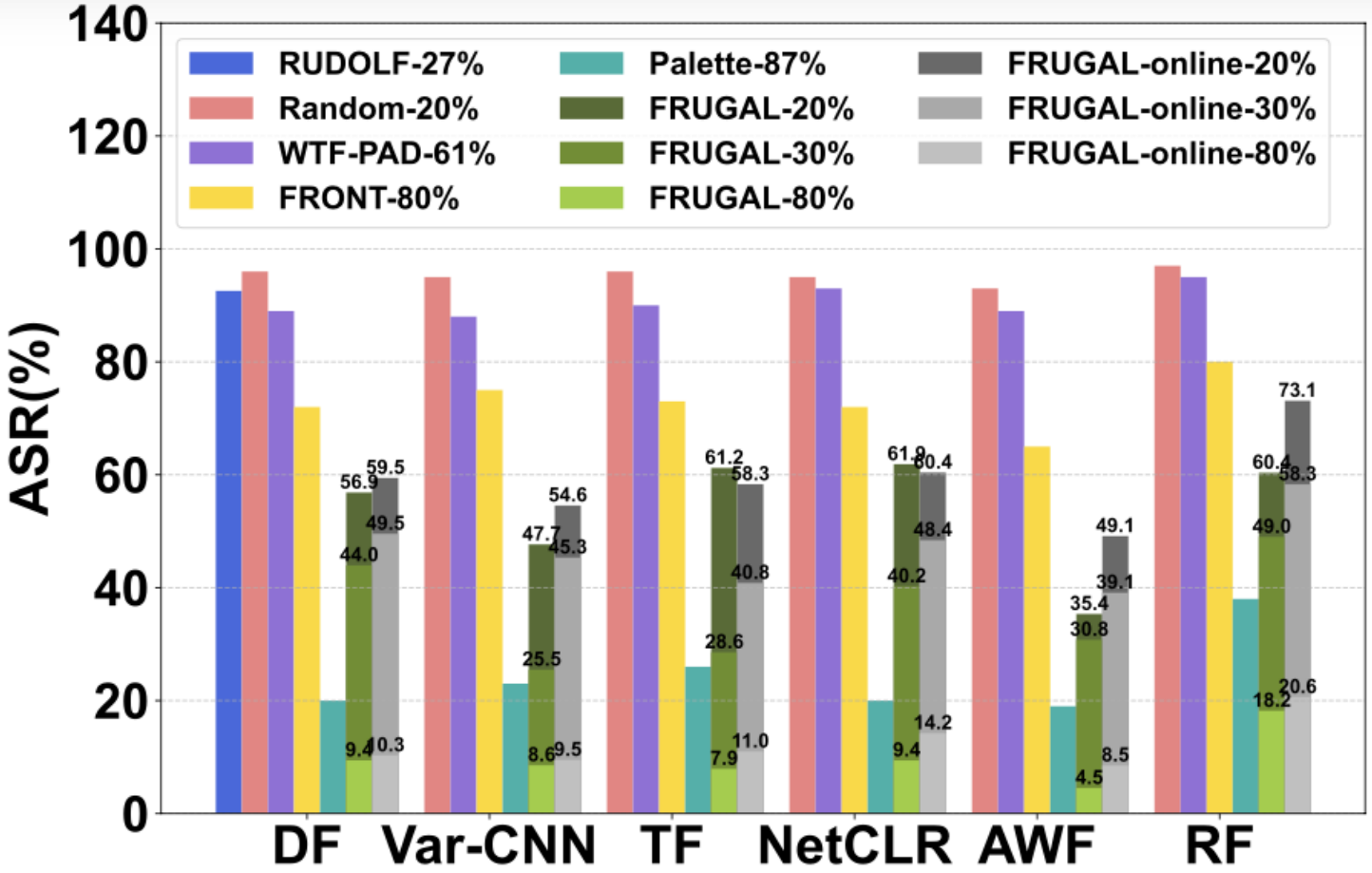
Result: Real-World Generalization

Adversarial training retrains models on defended traffic to "recapture" hidden signals.

Minimal Information Leakage: At 80% BWO, FRUGAL maintains ASR at **9.42%**, whereas Palette climbs to **20.27%**.



Result: Robustness Against Adversarial Training



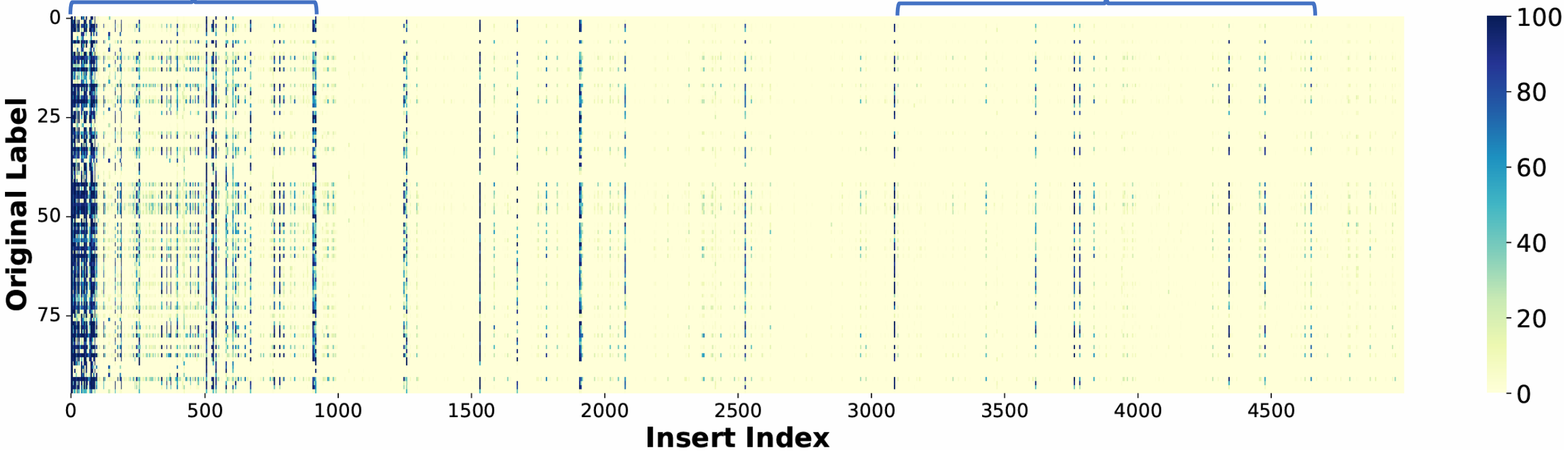
Both FRUGAL and FRUGAL-online can achieve SOTA adversarial training performance under similar BWO.

Strategic Insight

Visualizing the Learned Policy

FRUGAL learns to mask the initial phase, which contains the most identifying information.

Minimal injections required after the initial phase.



Conclusion: A New Benchmark for WFD

New Perspective:

Shifted optimization from simple classification error to Mutual Information Minimization.

Notable Efficiency:

Achieved <3% ASR with only 30% Bandwidth Overhead.

Adversarial Robustness:

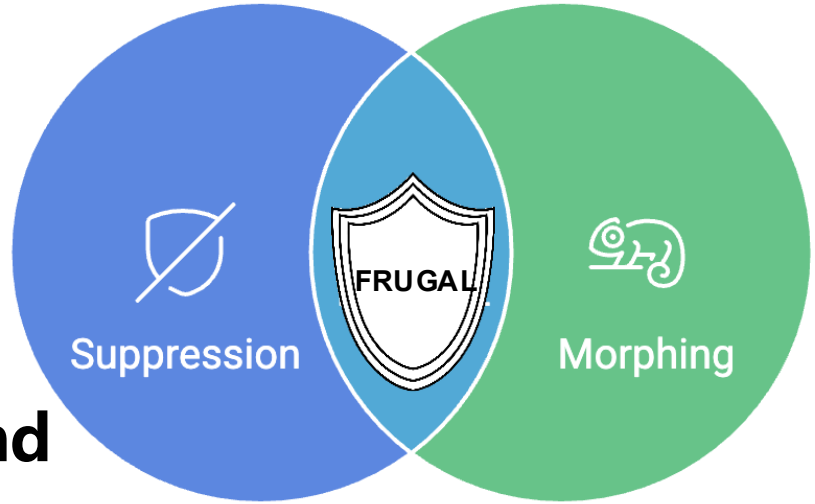
Solved 'Drift' via Dynamic Feature Elimination, securing against adaptive attackers.

Practicality:

FRUGAL-Online can prove real-time, low-latency deployment for actual users.

Conclusion: Cease at the Ultimate Goodness

Is FRUGAL a suppression-based or morphing-based defense?



Our Perspective: "Yes and No."

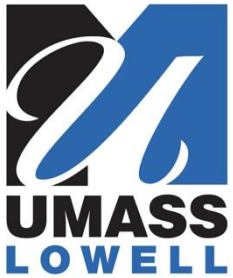
- Boundaries are blurred.
- Actual resource efficiency is more meaningful.

At what cost can we afford before we cease, and to what extent do we expect the goodness?

- FRUGAL provides the answer: Tell me the cost where you'd like to cease, and we will provide the ultimate goodness at that level.


SOUTHEAST UNIVERSITY
東南大學
1902
南京
心於至善


福耀科技大學
FUYAO UNIVERSITY OF SCIENCE AND TECHNOLOGY


UMASS
LOWELL

Thank you !