MICHIGAN STATE UNIVERSITY

OPTML

NEURAL INFORMATION PROCESSING SYSTEMS

# MSU OPTML Lab @ NeurIPS 2024
# Menu of Innovations

## Conference Papers | East Exhibit Hall A-C

### From Trojan Horses to Castle Walls: Unveiling Bilateral Data Poisoning Effects in Diffusion Models

**Chefs:** Zhuoshi Pan*, Yuguang Yao*, Gaowen Liu, Bingquan Shen, H. Vicky Zhao, Ramana Kompella, Sijia Liu (Equal contribution)

**Serving Time:** Wed (Dec. 11) 11:00-14:00, Poster #4602

**Key Ingredients:** Backdoor Attack, Diffusion Models

### UnlearnCanvas: Stylized Image Dataset for Enhanced Machine Unlearning Evaluation in Diffusion Models

**Chefs:** Yihua Zhang, Chongyu Fan, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Gaoyuan Zhang, Gaowen Liu, Ramana Kompella, Xiaoming Liu, Sijia Liu

**Serving Time:** Wed (Dec. 11) 16:30-19:30, Poster #4309

**Key Ingredients:** Machine Unlearning, Diffusion Models, Benchmark

### WAGLE: Strategic Weight Attribution for Effective and Modular Unlearning in Large Language Models

**Chefs:** Jinghan Jia, Jiancheng Liu, Yihua Zhang, Parikshit Ram, Nathalie Baracaldo, Sijia Liu

**Serving Time:** Thu (Dec. 12) 16:30-19:30, Poster #4300

**Key Ingredients:** Large Language Models, Modularity, Unlearning

### Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models

**Chefs:** Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, Sijia Liu

**Serving Time:** Fri (Dec. 13) 11:00-14:00 Poster #2509

**Key Ingredients:** Adversarial Unlearning, Diffusion Models

## Workshop Organization & Papers

### The 3rd New Frontiers In Adversarial Machine Learning

**Saturday (Dec. 14) | East Ballroom C**

### Adversarial Watermarking for Face Recognition

**Chefs:** Yuguang Yao, Anil Jain, Sijia Liu

**Serving Time:** Sat (Dec. 14), AdvML-Frontiers Workshop (East Ballroom C)

**Key Ingredients:** Watermarking, Adversarial Attacks, Biometrics

### Rethinking Negative Preference Optimization for LLM Unlearning

**Chefs:** Chongyu Fan*, Jiancheng Liu*, Licong Lin*, Jinghan Jia, Ruiqi Zhang, Song Mei, Sijia Liu (Equal contribution)

**Serving Time:** Sun (Dec. 15), SafeGenAI Workshop (Exhibition Hall A)

**Key Ingredients:** Preference Optimization, LLM Unlearning

## A Heartfelt Thank You to Our Incredible Collaborators

IBM Research · cisco · intel · DSO NATIONAL LABORATORIES · M · THE UNIVERSITY OF CALIFORNIA BERKELEY 1868 · 清華大學 Tsinghua University