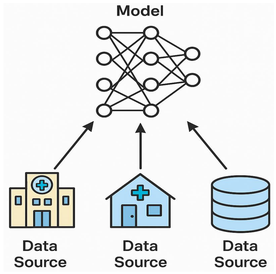
Adversarial Attacks against Closed-Source MLLMs via Feature Optimal Alignment

WASHINGTON STATE UNIVERSITY
TRI-CITIES

WASHINGTON STATE UNIVERSITY
TRI-CITIES

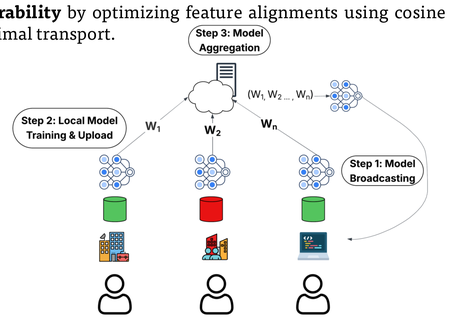
Introduction

❑ Closed-Source MLLMs



❑ Multimodal LLMs (MLLMs)

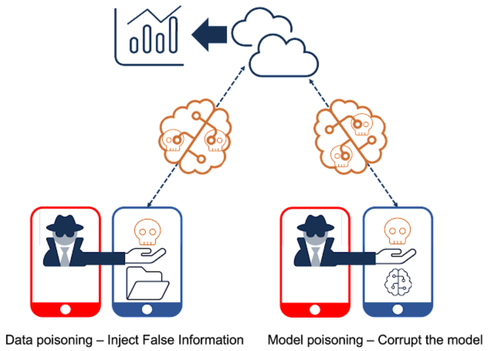
Multimodal Large Language Models (MLLMs) like GPT-4o, Claude, and Gemini are increasingly deployed in real-world applications. However, their **closed-source nature** poses significant challenges for adversarial robustness evaluation. FOA-Attack enhances **adversarial transferability** by optimizing feature alignments using cosine similarity and optimal transport.



Threat Model

Attacker Goals

- **Untargeted attacks**: Degrade overall model performance or prevent convergence.
- **Targeted attacks**: Manipulate the global model to misclassify specific inputs.



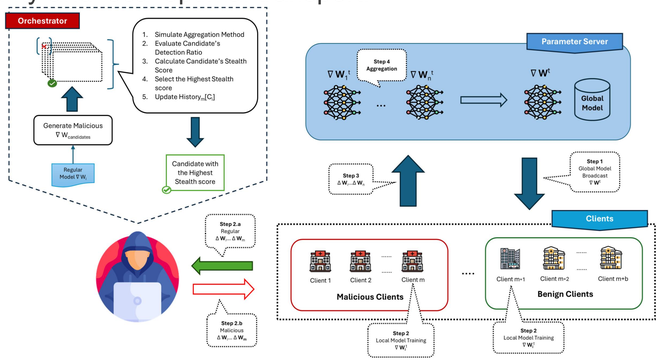
Attacker' capabilities

- FOA-Attack **operates in a strict black-box setting** with no access to target model parameters, gradients, or API queries.
- Only **open-source surrogate MLLMs** (LLaVA, InternVL2) are used to generate transferable adversarial examples targeting closed-source MLLMs.

FOA-Attack

We propose **FOA-Attack**, a novel black-box adversarial attack framework that improves transferability to closed-source MLLMs via Feature Optimal Alignment combining global cosine similarity and local optimal transport.

1. Attack Overview



Feature Extraction: Compute patch-level token features from the surrogate model for both the adversarial image and the target reference.

Malicious Gradient Collection: Solve the optimal transport plan to match local feature distributions between surrogate and black-box target representations.

Gradient Modification by OS: Update adversarial perturbation by minimizing both global cosine similarity loss and local OT-based alignment loss jointly.

Coordinates Malicious Clients: Combine global feature alignment (cosine similarity) and local feature alignment (optimal transport) into unified FOA loss.

Tailors Poisoning via Gaussian Sampling: Apply PGD-based perturbation update bounded by L-inf norm constraint ε=16/255 for each iteration until convergence to produce the final adversarial example.

Evaluates Stealth with Custom Metrics: FOA-Attack achieves state-of-the-art attack success rates on GPT-4o, Claude-3, and Gemini, significantly outperforming baselines like M-Attack, CroPA, and AdvCLIP.

Optimizes for Evasion & Impact: OT-based local alignment captures fine-grained structural correspondences that global cosine similarity alone cannot capture, boosting transferability.

Final Gradient Submission: The final adversarial image fools closed-source MLLMs into generating targeted harmful outputs while remaining imperceptible to human observers.

2. Global Feature Alignment

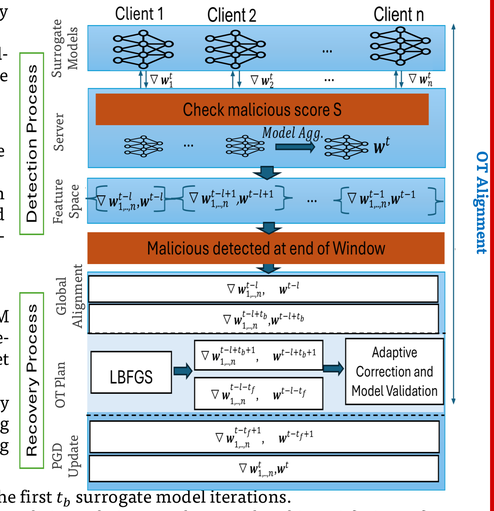
Alignment Goal: Align adversarial image features globally with the target output representation across multiple surrogate MLLMs.

Detection Methodology: Maximize cosine similarity between adversarial image features and target text embeddings in the shared multimodal feature space of the surrogate model, promoting feature-level alignment without access to the black-box target.

3. Local Feature Alignment

Challenge : Local feature distributions differ across MLLM architectures, making direct cosine similarity insufficient for fine-grained patch-level alignment between surrogate and black-box target visual encoders.

Recovery Methodology: we Propose A window-based recovery scheme restores global model without full retraining, reducing computational overhead and enhancing robustness against recurring attacks using three steps.



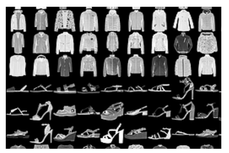
- **Step 1 - Feature Extraction:** PS collects exact client updates for the first t_b surrogate model iterations.
- **Step 2 - OT Alignment:** Compute OT plan minimizing Wasserstein distance between adversarial and target features for t_c fine-grained patch-level alignment.
- **Step 3 - PGD Update:** Apply PGD update minimizing FOA loss (global cosine + local OT) within ε=16/255 budget for t_f iterations until convergence.

Experimental Setup

❑ Evaluation benchmarks: AdvBench and SafeBench.
- Surrogate models: LLaVA-1.5, InternVL2, MiniGPT-4

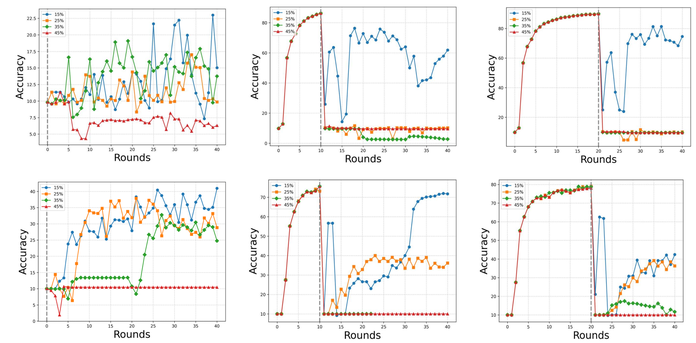
❑ Target black-box MLLMs: GPT-4o, Claude-3-Opus, Gemini-1.0-Pro. Perturbation budget: ε=16/255, step=2/255, 200 PGD iterations.

❑ We evaluated our proposed attack in two experimental settings:
- Single-surrogate transfer: adversarial examples crafted on one surrogate, evaluated on all black-box targets.
- Ensemble-surrogate transfer: multiple surrogates combined to further improve cross-model transferability to GPT-4o, Claude-3, and Gemini.



Results

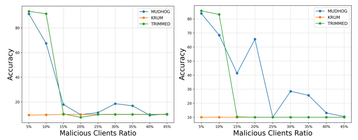
❑ *Accuracy Degradation Resulting from Our Attack Initiated at Different Rounds (0,10,20) on AdvBench (First Row) and SafeBench (Second Row) Datasets*



❑ *Attack Success Rate (%) Across Different Surrogate Models and Target MLLMs*

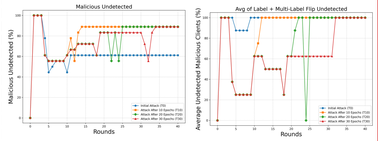
Malicious Ratio	100 Clients				200 Clients				300 Clients			
	MNIST		Fashion MNIST		MNIST		Fashion MNIST		MNIST		Fashion MNIST	
	Acc. (%)	Err. Rate (%)	Acc. (%)	Err. Rate (%)	Acc. (%)	Err. Rate (%)	Acc. (%)	Err. Rate (%)	Acc. (%)	Err. Rate (%)	Acc. (%)	Err. Rate (%)
10%	9.82	90.18	36.07	63.93	9.82	90.18	10.0	90.0	9.8	90.2	10.0	90.0
20%	10.09	89.91	10.0	90.0	9.8	90.2	10.0	90.0	9.8	90.2	10.0	90.0
30%	9.8	90.2	10.0	90.0	9.8	90.2	10.0	90.0	9.58	90.42	9.99	90.01
40%	10.32	89.68	10.0	90.0	9.8	90.2	10.0	90.0	10.1	89.9	10.0	90.0
50%	9.8	90.2	10.0	90.0	9.8	90.2	10.0	90.0	9.8	90.2	10.0	90.0

❑ *Impact of Our Attack on MUD-HoG [3], Trimmed Mean[4], and Krum[5] Across Different Malicious Client Ratios*



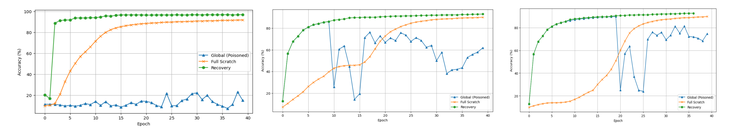
MNIST *Fashion-MNIST*

❑ *Attack Success Rate vs. Perturbation Budget ε on GPT-4o and Claude-3*



FOA-Attack *M-Attack (Baseline)*

❑ *Accuracy Comparison of Our Defense + Recovery Mechanisms vs. Full Training from Scratch Across Different Windows*



References

[1] **M. Shaaban**, **A. Abdelnaby**, and **Jia, X. Jia et al.**, "Adversarial Attacks against Closed-Source MLLMs via Feature Optimal Alignment", under review for the 30th European Symposium on Research in Computer Security (ESORICS), 2025.
[2] Zhao, Lavaur, et al. BusnelM-Attack, Autrel. Systematic analysis of label-flipping attacks against federated learning in collaborative intrusion detection systems. In Proceedings of the 19th International Conference on Availability, Reliability and Security, pages 1–12, 2024
[3] Luo et al. CroPA: Cross-prompt adversarial attack on vision-language models. ICLR 2024.
[4] Zhang, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In International conference on machine learning, pages5650–5659. Pmlr, 2018
[5] Villani, Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. Advances in neural information processing systems,30, 2017.

Code & Paper

SCAN ME