Robust and Efficient Multimodal Intelligence

Overview

This project investigates the fundamental mechanisms behind how multimodal models synthesize and interact with diverse input streams, including audio, text, and vision. While current AI systems show remarkable capabilities, the "black box" nature of diverse modality fusion and the overall reasoning requires a deeper investigation. This research aims to: (1) **Interpret** the model internals to understand how semantic information is shared between modalities; (2) **Enhance model robustness** against diverse multimodal adversarial attacks; and (3) **Develop computationally efficient algorithms** leveraging the learnings from model interpretability. By stress-testing multimodal models in security-critical applications, the project ensures that next-generation AI systems are safe, resilient and computationally sustainable.

Intellectual Merit

This project advances the field of Artificial Intelligence by providing a granular understanding and efficient solutions for cross-modal semantic interactions and vulnerabilities.

- **Interpretable Fusion:** We will develop novel white-box tools to map how information flows between modality-specific encoders and the joint latent space.
- **Robustness:** We will introduce new learning paradigms that explicitly penalize cross-modal leakage during attacks, creating a "safety-by-design" framework.
- **Efficiency:** We will drive **(a) task-aligned efficiency**, where computational power is shifted in real-time to the most relevant modality critical for a given task, and **(b) semantic efficiency,** where redundancies within a given modality are dynamically leveraged without sacrificing generative or reasoning capabilities.

Broader Impacts

The societal implications of this work address the urgent need for secure and accessible AI.

- **National Security:** By stress-testing models in safety-critical domains, the project provides a path forward for deploying AI in applications where reliability and safety is non-negotiable.
- **Environmental Sustainability:** The proposed efficiency measures reduce the massive energy footprint of large-scale generative models, making AI more sustainable and accessible to researchers with limited hardware.
- **Education and Inclusion:** This project will integrate research findings into undergraduate and graduate curricula, fostering a diverse workforce trained in AI safety and interpretability.