Shreyansh Modi

shreyansh_m@ee.iitr.ac.in +91-8287512050 Shreyansh Modi shrrynsh Shreyansh Modi

Education

St. Teresa School *2010 – 2024*

Indian Institute of Technology, Roorkee *2024 – 2028*
B.Tech in Electrical Engineering

Experience

Machine Learning Engineer Intern | *Infigro.ai* *July 2025 – Nov 2025*

- Built a Data Analysis Agent capable of interpreting Bar Charts and Trendline Data for automated insights generation.
- Worked extensively on Data Interpretation and Analysis pipelines to extract meaningful patterns from visual data representations.

Research Intern| *Electrical Engineering Department, IIT Roorkee* *Dec 2024 – Jan 2025*

- Research focused on the intersection of Control Systems and Cybersecurity, advanced control strategies to enhance the security and resilience of cyber-physical systems.
- Worked on Arduino,Matlab to simulate various types of attacks on cyber-physical systems.
- Devised new strategies to mitigate different kinds of cyber attacks on cyber-physical systems.

Publications

Guidance for Low-Level Perceptual Editing in Unconditional Diffusion Models

S. Modi, A. Tomar, A. Aggarwal arXiv | Code *CVPR'26 GCV Workshop*

A method that enables semantic control over unconditional diffusion models (DDPMs) by extracting concept-specific direction vectors from the model's internal h-space (the bottleneck activation of the U-Net architecture) and applying them as targeted perturbations during the reverse generative process.

How Many Counterfactuals Does It Take? Probing VLM Hallucinations Through Circuits and Causal Effects

S. Modi, A. Tomar, A. Gupta, S. Singh, A. Sinha arXiv | Code *arXiv Preprint, 2026*

A study of the sample complexity of counterfactual robustness for hallucinated outputs in Visual Language Models, introducing a causal influence metric and using circuit discovery to derive empirical bounds on the counterfactual samples needed to reliably detect instability in hallucinated predictions.

Projects

Adobe Photo Editor *Nov 2025 – Dec 2025*

Developed a photo editing application featuring automated editing capabilities powered by Computer Vision techniques. Implemented intelligent image processing pipelines enabling context-aware, automated photo enhancements and edits without manual intervention. Secured **4th place** in **InterIIT Tech Meet 14.0**

VLM Hallucination Mitigation via Self-Verification Decoding *Jun 2025 – Sep 2025*

Developed an optimization framework to mitigate visual hallucinations in multimodal Vision-Language Models (VLMs). Implemented decoding-time interventions to ensure generated text remains strictly grounded to image inputs. Evaluated and improved model reliability using standardized benchmarks like POPE and CHAIR, significantly enhancing the semantic alignment and trustworthiness of the model's multimodal outputs.

InVisionDX: CNN-Based Disease Detection Model *Mar 2025 – Apr 2025*

Developed Web2-integrated disease prediction models based on Computer Vision and Convolutional Neural Networks, achieving over 95% accuracy. Incorporated several CNN-based models for predicting diseases such as COVID-19, Pneumonia, Tuberculosis, Lung Cancer, and Alzheimer's disease from visual data.

Rethinking-CD: A Reproducibility Study on the Effectiveness of Contrastive Decoding *Jun 2026 – Aug 2026*

Conducted a reproducibility study critiquing contrastive decoding methods used to fix object hallucinations in Multimodal LLMs. Evaluated these techniques on LLaVA-1.5 using the POPE benchmark, ultimately showing that their supposed performance gains do not come from better visual grounding; instead, they are mathematical side-effects of distribution shifts and adaptive plausibility constraints during decoding.

RE-UltraBreak: Reproducibility Study of Universal and Transferable Jailbreak Attacks on Vision-Language Models *Jun 2026 – Aug 2026*

Conducted a reproducibility study on the paper UltraBreak to audit safety vulnerabilities in Vision-Language Models. Replicated the optimization pipeline to generate universal, cross-model adversarial image jailbreaks using a single surrogate model. Validated that randomized image transformations and semantic loss relaxations smooth the loss landscape, preventing overfitting and exposing critical transferable security flaws across black-box VLMs.

Key Courses Taken

Mathematics: Linear Algebra, Probability and Statistics,Numerical Methods, Optimization,Calculus
Other Relevant Courses: C++ Programming,Machine Learning, Deep Learning

Positions of Responsibility

Core Member: Data Science Group , IIT Roorkee *Jan 2025 - Present*