Pranav Gupta

+1(561)-229-4245 gpranav@umich.edu pranavgupta2603.github.io pranavgupta2003 pranavgupta2603 Scholar

Education

- **University of Michigan**, Ann Arbor (ongoing) Aug. 2025 – May. 2027
  MS (Computer Science) - **GPA**: 3.98
  Courses: *Advanced Scalable Systems in Agentic AI, Advanced Computer Vision, Natural Language Processing, Action and Perception, Graphics and GenModels, Building Small Language Models*
- **SRM Institute of Science and Technology**, Chennai, Tamil Nadu Aug. 2021 – Aug. 2025
  B.Tech (Computer Science)

Publications

- *Fulcrum: Power and Latency-aware Optimization of Concurrent DNN Training and Inference on Edge*, ACM Middleware '26
- *Pagoda: Roofline Characterization of Energy and Time for DNN Inference and Training on Edge Accelerators*, ICPP 2026
- *ViDAS - Vision-based Danger Assessment and Scoring*, ICVGIP 2024.
- *ECHO: Environmental Sound Classification with Hierarchical Ontology-Guided Semi-Supervised Learning*, IEEE CONECCT 2024.

Work Experience

Research Assistant, University of Michigan, **PI** - Dr. Jeong Joon Park Sept. 2025 – Present

- Built a cascading **flow** model that generates novel **surface fields** to extract high-quality and **topologically robust** meshes.
- Engineered a field-conditioned flow model to produce controllable generation of meshes, submitted to CVPR 2027.
- Trained **masked auto-encoders** and **LeJEPA** to learn view-invariant embeddings for multispectral subsets of hyperspectral images.
- Classified **hyperspectral** images in a novel dataset of different chemical solutions and concentrations in spillages.

Research Intern, Stanford University, **PI** - Dr. Teddy Akiki Jun. 2024 – Jun. 2025

- Developed autoregressive and diffusion models (**Chronos, TimeLLM**) for time-series forecasting to uncover latent "language-like" patterns in **BOLD fMRI** signals.
- Designed and trained a **VQ-VAE + GPT2/LLaMA2** pipeline to discretize continuous fMRI signals into latent tokens and predict future sequences.

Software Developer Intern, Kiwi - Repository Apr. 2025 – May. 2025

- Constructed a **LangGraph multi-agent framework** handling **100+** customer service **interactions per day** across varied query types like Transactions, Rewards, and Card applications.
- Framework included **15+ tools** and Freshdesk and Mixpanel integrations, delivering accurate results in $<$ 1min.

Research Intern, Indian Institute of Science, **PI** - Dr. Yogesh Simmhan Aug. 2024 – Feb. 2025

- Proposed and improved baseline neural networks through a **penalizable MAPE loss function** in optimizing concurrent training and inference on edge accelerators by **80%**.
- Built a **novel Active Learning technique** that samples the best power modes on Jetson devices for DNN workloads with errors close to optimal (0%).

Research Intern, Samsung R&D, **PI** - Mr. Shwetank Choudhary Sept. 2023 – Jul. 2024

- Devised a **novel** two-stage pipeline that trains a CNN with semi-supervised learning(ResNet50, EfficientNet) on the label ontology of environmental sound classes as a **pretext** task.
- Produced an accuracy **increase** of 1-8% compared to baseline models, achieving 97.5% **state-of-the-art** accuracy in the ESC-10 benchmark dataset.

Projects

Aether - *vLLM, Knowledge Graphs, Temporal-Aware Retrieval, LLM Evaluation* Apr. 2026

- Developed a hypergraph **memory pipeline** for LLMs with typed extraction and temporal retrieval, **outperforming** baselines by **50.8%** in LLM-judge accuracy and **reduced** response latency by **60.8%** compared to Mem0.

MedScribe - *OpenAI-SDK, Clinical NLP, Summarization, Structured Outputs, Biomedical Models* Dec. 2025

- Engineered a **multi-agent** pipeline converting **clinical conversations into FHIR records**, developing the GPT-5 baseline and evaluation framework that **demonstrated 9.7% higher** perfomance than GPT-5.

Skills

- **Languages:** Python, C++, SQL
- **ML / Deep Learning:** PyTorch, TensorFlow, Semi-supervised learning, Self-Supervised learning
- **Computer Vision / Generative AI:** Diffusion Models, Vision-Language Models, Hyperspectral Imaging, 3D Mesh Generation, Voxel Representations, Agentic-AI SDK, RAG, LangGraph, vLLM
- **3D Vision / Graphics:** 3D Mesh Generation, Light Physics, Geometric Mathematics
- **HPC / Systems:** SLURM, GPU Computing, Distributed Training, Linux, HPC Clusters