

# VAISHNAV POTLAPALLI

☎ 332-373-5002

✉ [vp2499@nyu.edu](mailto:vp2499@nyu.edu)

🌐 [linkedin.com/in/vaishnav-potlapalli](https://www.linkedin.com/in/vaishnav-potlapalli)

🐙 [github.com/vaishnav9v](https://github.com/vaishnav9v)

## Publications

### PromptIR: Prompting for All-in-One Blind Image Restoration

NeurIPS 2023

*Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan*

- Proposed an implicit prompt-learning based approach for All-in-One blind Image Restoration. Achieves SoTA performance on multiple image restoration tasks, without any prior degradation information.

### Sketch3T: Test-Time Training for Zero-Shot SBIR

CVPR 2022

*Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, Yi-Zhe Song*

- Introduced a novel test-time training paradigm for zero-shot sketch-based image retrieval that adapts to new categories and sketch distributions using a single sketch, outperforming state-of-the-art methods.

### MediTables IIIT

GREC 2021

*Akshay Praveen Deshpande, Vaishnav Potlapalli, Ravi Kiran Sarvadevabhatla*

- Built a new dataset and semantic segmentation model for camera captured medical document images.

## Experience

### Sedai Inc

March 2025 – Present

*Software Engineer - AI/ML*

- Designed an **autonomous optimization engine** for Databricks using heuristic search + time-series forecasting to recommend right-sizing, scheduling, and safe auto-termination.
- Implemented a **rule-based LLM router** (allow/deny lists, default fallbacks, per-project policies, cost/latency budgets) with health- and error-aware failover across OpenAI/Anthropic/Bedrock backends.

### Floma Inc - AI Agents for Marketing

January 2025 – March 2025

*Founding Software Engineer - Machine Learning*

- Engineered a CV auto-labeling pipeline fusing OW-Detection (Grounding-DINO) with SAM segmentation, achieving a **5×** dataset-annotation speedup.
- Fine-tuned Llama-family models on ad-copy data via **SFT /RLHF**, boosting generation performance by **23%**.
- Built a **multimodal LLM agent** that programmatically composes dynamic display ads; created an SVG generation/editing module for size-agnostic ad rendering and on-the-fly customization.

### MBZ University of Artificial Intelligence

July 2022 – July 2023

*Research Assistant - Computer Vision (Advisor: Dr. Salman Khan)*

- Proposed and implemented a novel Visual transformer based prompt-learning framework for All-in-one blind Image Restoration / Enhancement called **PromptIR**, which achieved **SoTA** performance improving over previous methods by **0.9 dB** on dehazing, deraining and denoising benchmarks. Work presented as part of **Neurips 2023**
- Adapted computer vision based continual learning techniques **L2P**, **DualPrompt** methods for video action recognition improving performance over previous techniques by over **10%** accuracy and **14% BWF**, on several public benchmarks.
- Studied parameter-efficient finetuning techniques to improve downstream performance of **Multimodal LLM models**.

## Education

### New York University Courant Institute of Mathematical Sciences

Sep. 2023 – Dec 2024

*Masters of Science in Computing, Entrepreneurship, and Innovation GPA: 3.96/4.0*

*New York City, NY*

*Relevant Courses: LLVMs, Big Data and ML Systems, Foundations of Computer Networks*

*Honors: M. Michael Waller Master's Fellowship*

## Projects

### Efficient Mixture-of-Depths(MoD) LLM Inference | PyTorch, CUDA, LLMs

March 2024 – May 2024

- Engineered a Mixture-of-Depths (MoD) transformer on a LLaMa-style baseline (55M parameters, 6 layers) by integrating dynamic token routing with top-k selection, auxiliary loss, and an auxiliary MLP predictor
- Utilized Torch CUDA events for profiling and simulated a novel GPU scheduling policy to boost throughput, lower latency, and validate improvements via ablation studies and perplexity analysis.

### PITCHPAL: AI-Powered Presentation Coach | FastAPI, React, LLMs, AI Agents October 2024 – December 2024

- Implemented a multi-modal AI Agent with a fast-api backend that combined advanced speech recognition, natural language processing, and computer vision to analyze presentation content and delivery and provide relevant feedback.

## Technical Skills

**Languages/Frameworks:** Python, Java, C/C++, CUDA, TypeScript, Pytorch, TF, OpenCV, Transformers, Django, MySQL