SUMAN MADIPEDDI

smadiped@asu.edu | (602) 565-9192 | linkedin.com/in/suman-madipeddi |GitHub |Portfolio

SUMMARY

AI Engineer with **2+ years** of experience building production-grade systems in Generative AI, LLMs, and Computer Vision. Skilled in deploying scalable ML pipelines, fine-tuning, and optimizing inference. Passionate about applying research to real-world AI products.

EDUCATION

M.S. Robotics and Autonomous Systems **Expected May 2025**
Arizona State University, Tempe, Arizona 3.74 GPA

- **Relevant Coursework:** Artificial Intelligence, Applied AI and ML, Reinforcement Learning and Perception in Robotics

TECHNICAL SKILLS

Deep Learning: PyTorch, TensorFlow, CoreML, TensorRT, Keras, CNN, RNNs, GANs, Computer Vision, Multimodals (CLIP, BLIP)

AI/ML & GenAI: Transformers, LLMs (GPT, Gemini, Claude), Stable Diffusion models, JAX, VLMs, HuggingFace, LangChain, LlamaIndex, RAG, FAISS, Pinecone, Conversational Agents, Fine-tuning, RLHF, OSS ML models, AGI, OpenAI, Twilio

Languages & Frameworks: Python, C, C++, Go, Swift, Java, TypeScript, Django, Node.js, Next.js, React JS, React Native, CUDA

Data & MLOps: Pandas, NumPy, ETL, SQL, GPU, PostgreSQL, Redis, Databricks, Amazon (SageMaker, Bedrock), CI/CD, kafka

Cloud & DevOps: AWS (S3, EC2, Lambda), GCP, Azure, Docker, Nvidia Triton, Linux, Snowflake, Agile, Apache Spark, REST APIs, gRPC

PROFESSIONAL EXPERIENCE

Minor Chores, USA: AI and LLMs Engineer **Jan 2025 – Present**

- Designed and deployed a **RAG-based conversational AI system** (Vertex AI, RLHF) with multimodal inputs, achieving 90% intent accuracy, and built an **LLM-based recommendation engine** leveraging geospatial + behavioral data for real-time chore-preneurs
- **Modernized UI/UX across iOS/Android** by leading React Native, Swift, and TypeScript development within a microservices architecture; orchestrated deployment with Kubernetes, ensuring scalable, zero-downtime rollouts.
- **Improved performance and engagement** by reducing latency 40%, introducing broadcast + interactive messaging, and driving measurable gains in onboarding 25% and user engagement 30%.

PROJECTS

Voice AI Agent for Automated Lead Qualification – Setter.AI (Twilio, Deepgram, Typescript, Docker) **Aug 2025**

- **Engineered a Node.js (TypeScript) + Twilio** outbound calling platform, integrating with Go High Level CRM to automate lead engagement within 10 minutes of creation, orchestrated via Docker for scalable reliability.
- **Built a GPT-4 powered conversational AI pipeline** with Deepgram STT/TTS for natural lead screening, appointment scheduling, and real-time CRM data logging through a React/TypeScript dashboard.
- **Delivered 24/7 automation** that eliminated 25+ hours of manual work per week, accelerated response times, and established a scalable foundation for client growth.

Multi-Modal Text-to-Video System **Aug 2025**

- Designed and deployed a **multi-modal generative pipeline** using **Stable Diffusion v1.5, and Wan 2.2-T2V-A14B**, producing **10+ minute 720p/24fps videos** with synchronized audio and high-quality images from text prompts.
- Built a **React frontend and Express.js backend**, integrating **FFmpeg, PyTorch, Hugging Face Diffusers,** with low-latency media generation by scenes and then stitching all the scenes and generating the video workflows.
- Optimized GPU utilization with **80GB+ VRAM, distributed inference, mixed precision, and frame interpolation (RIFE)**, reducing rendering latency by **40%** and enabling **scalable multi-hour content generation**.

On-Device Real-Time Gesture Recognition System **June 2025**

- Developed a real-time, on-device gesture and voice recognition system that dynamically overlays emojis and confetti effects on live video streams, enhancing user interaction for video calls and streaming platforms.
- Trained gesture classification models with **Flax/JAX**, optimized the inference pipeline with **ONNX Runtime** and **CUDA,** achieving low-latency inference and robust performance across **17** hand gestures. Integrated **Whisper**-based voice command recognition for spontaneous, conversational emoji and effect triggers, supporting multimodal interaction.
- Architected a **modular, privacy-first overlay pipeline** using OpenCV and Pillow enabling integration with **FaceTime and OBS.**

Object Segmentation on ARMBench (PyTorch, R-CNN, ResNet-50) **May 2024**

- Developed a Mask R-CNN model with a ResNet-50 backbone in PyTorch for object segmentation on the ARMBench dataset (50K+ images, 450K+ segments).
- Trained across three subsets (*mix-object*, *zoomed-out*, *same-object*) to evaluate performance in occluded and cluttered scenes.
- Achieved **mAP@50** scores of 0.48, 0.41, 0.48 and **mAP@75** scores of 0.48, 0.06, 0.38 across subsets, using COCO-style evaluation and OpenCV visualizations to monitor segmentation accuracy and generalization and tested on other data ingestion.