# SAMEEP VANI

svani@asu.edu | Linkedin | GitHub | Google Scholars | +1 602-561-0176 | San Francisco Bay Area, CA

## EDUCATION

**Arizona State University** Tempe, AZ
*Master of Science in Computer Science* 08/2023 - 12/2025
**Focus:** Computer Vision, Natural Language Processing (NLP), Video Large Language Models (LLMs)

**Ahmedabad University** Ahmedabad, India
*Bachelor of Technology in Computer Science* 07/2019 - 06/2023
**Focus:** Machine Learning, Computer Vision, Natural Language Processing (NLP)

## TECHNICAL SKILLSET

**Programming Languages:** Python, R, C/C++, Java, JavaScript, SQL, HTML/CSS, MATLAB
**ML & Data Tools:** PyTorch, TensorFlow, HuggingFace, Keras, Scikit-learn, NLTK, LangChain, Tableau, Browser-Use, LangGraph
**Web & API Development:** Django, Flask, FastAPI, Node.js, Express.js, React.js, Next.js, REST, GraphQL, Bootstrap, TailwindCSS
**Cloud & DevOps:** AWS, GCP, Azure, Docker, Kubernetes, Nginx, Load Balancer
**Distributed Systems:** Kafka, Redis, Celery, Apache Airflow, Apache Spark, Hadoop
**Databases:** MySQL, PostgreSQL, MongoDB, HBase, Firebase, Neo4j, VectorDB
**Methodologies:** Chain-of-Thought, RAG, Direct Preference Optimization, In-Context Learning, Attention Mechanism

## EXPERIENCE

**Sphinx Labs, Inc.** San Francisco, CA
*Founding AI Engineer Intern [Python, Docker, Django, Nodejs]* 05/2025 - 08/2025

- Led the development as a founding engineer at a YC-backed startup, architecting a core compliance automation system.
- Engineered browser agents for client data retrieval, reaching a 94% success rate across multiple compliance platforms.
- Designed Django models with PostgreSQL for structured storage and processing of case alerts and document checks.
- Optimized media research pipelines via prompt engineering, improving retrieval performance and tailoring for KYB/KYC/AML decisioning.
- Integrated Checkin API for automated ID and proof-of-address verification, reducing compliance processing time by 42%.

**Arizona State University - Guided by 'YZ' Yezhou Yang** Tempe, AZ
*Machine Learning / Computer Vision / NLP [Python, Pytorch, HuggingFace]* 05/2024 - 05/2025

- Identified gaps in current video understanding benchmarks, focusing on overall description rather than temporal aspects.
- Evaluated 1,000+ video samples across 4 benchmarks, identifying a 28% gap in temporal understanding for top Video-LLMs.
- Introduced TimeWar, a benchmark with annotated sequences, revealing a 32% drop in model accuracy on temporal tasks.
- Built a data pipeline generating 10k+ preference pairs for DPO fine-tuning, boosting test performance by 9%.
- Bridged the gap of 7% between SOTA performance and baselines on temporal benchmarks and submitted to WACV 2026.

**Sculptsoft** Ahmedabad, India
*Machine Learning Intern [Python, Pytorch, Tensorflow, Reactjs, FastAPI, NLTK]* 01/2023 - 05/2023

- Executed diverse projects encompassing Churn Prediction utilizing Python, Movie Recommendations employing cosine similarity, and Fashion Image Classification leveraging a Big Data source, and achieved an accuracy of 98%.
- Created a Multimodal fusion-based model for classifying hateful memes, incorporating text and vision models via Visual BERT, Vision Transformer, and others, achieving an accuracy of 55% post-employing various optimization techniques.

## PROJECTS

**Video-Guided Instruction Retrieval & Answering** San Francisco, CA
*Software Engineer / AI Engineer / RAG [Python, Pytorch, HuggingFace]* 08/2025 - Present

- Developed an end-to-end multimodal RAG pipeline, combining Whisper-extracted transcripts and CLIP-encoded frames with a vector database for low-latency video question answering.
- Implementing Visual Chain of Thought, generated answers with frame-level attributions, enhancing interpretability.
- Fine-tuning using PEFT (LoRA) on Vicuna-7B to boost answer coherence and a lightweight temporal transformer adapter for accurate instruction sequencing.
- Engineering production infrastructure using FastAPI, Docker-Compose, Celery/Redis task queues, and Sentry monitoring.

**Symbolic Regression via ODEFormer** Tempe, AZ
*Representation Learning / Dynamical Systems [Python, Pytorch, HuggingFace]* 05/2024 - 12/2024

- Ran experiments comparing MLP and KAN architectures, improving regression accuracy by 12% on Lorenz-63.
- Replicated ODEFormer results on 100,000+ data points with transformer models for the symbolic regression task.
- Verified the use of transformer-based architectures to replace Neural ODEs for fitting time-dependent physical systems.