

SHREEJITH S G

+1(858)-518-9250 | ssuthrayegokulnath@ucsd.edu | [linkedin](#) | shreejithsg.github.io | [Google Scholar](#)

EDUCATION

University of California, San Diego - M.S. Data Science - GPA: 3.9	<i>Sep 2024 – Mar 2026</i>
Indian Institute of Technology, Madras - Diploma in Data Science- GPA: 4.0	<i>Apr 2023 – Jul 2024</i>
Vellore Institute of Technology, Chennai - B.Tech in CSE - GPA: 3.96	<i>Sep 2020 – Jul 2024</i>

TECHNICAL SKILLS

ML/DL Frameworks	PyTorch, CUDA, TensorFlow, Transformers (HuggingFace), DeepSpeed, FSDP, Ray, vLLM, SGLang, FlashAttention, Diffusion, Multi-modal Models, JAX
LLMs & GenAI	Fine-tuning (LoRA, PEFT, SFT), RLHF, DPO, Post-training Alignment, Model Compression, Distillation, Quantization, MoE, Multi-GPU Training, 3D Parallelism
Programming & Tools	Python, C/C++, SQL, Git, Docker, Kubernetes, AWS, Weights & Biases
Data Science	Causal Inference, Statistical Modeling, Spark, Dask, ETL Pipelines

EXPERIENCE

FastVideo / Hao AI Lab, UC San Diego	<i>Oct 2025 – Present</i>
<i>Open-Source Research Engineer – Diffusion Systems</i>	<i>California, US</i>
<ul style="list-style-type: none">Leading open-source contributor to FastVideo (3K+ GitHub stars), unifying 10B+ parameter video diffusion models with distributed training/inference infrastructure supporting multi-node H100 clustersDesigned and implemented novel LoRA extraction pipeline for 5B-parameter T2V diffusion checkpoints, enabling modular weight decomposition and 60%+ memory reduction while preserving generation qualityPorted and optimized MatrixGame-2 (world model) inference through hybrid attention refactoring (FlashAttention, sparse patterns, tiled computation), achieving 2.5x faster generation on H100 clustersCollaborating on research exploration for diffusion language models (dLLM), investigating architectural optimizations (KV-caching, sparse attention, semi-autoregressive mechanisms) for efficient text generation	
Scale AI	<i>Jun 2025 – Present</i>
<i>GenAI Specialist – Human Frontier Collective</i>	<i>California, US</i>
<ul style="list-style-type: none">Designed chain-of-thought evaluation frameworks for frontier reasoning models, creating structured rubrics across 15+ research benchmarks to assess factuality, logical coherence, and multi-step reasoning qualityContributed to post-training evaluation research for LLM alignment, analyzing RLHF trajectories across 500+ model checkpoints to surface failure modes and inform improvements in instruction-following and truthfulness	
Climate Analytics Lab, UC San Diego	<i>Apr 2025 – Present</i>
<i>ML Graduate Student Researcher – Advisor: Prof. Duncan Watson-Parris</i>	<i>California, US</i>
<ul style="list-style-type: none">Applied symbolic regression (PySR, KAN) to uncover non-linear relationships in climate data that black-box neural nets failed to capture, improving model interpretability & discovering new drivers of cloud variabilityBuilt diffusion-based probabilistic forecasting models (DDPM, EDM) for atmospheric variable prediction from satellite inputs, achieving 25% lower RMSE than physics-based baselines	

SELECTED PUBLICATIONS

BirdCLEF+ 2025 — Scalable Multi-Model Training & Knowledge Distillation CLEF 2025, 1st Author	
<ul style="list-style-type: none">Designed efficient one-vs-rest ensemble training 206 specialized classifiers in parallel, implementing per-species knowledge distillation to achieve > 0.90 mean AUCDeveloped targeted augmentation strategies, improving F1 from 0.00 to 0.72 for low-resource species	
NeuroFraudGAN — Synthetic data generation via GANs In Review, 1st Author	
<ul style="list-style-type: none">Collaborated with JPMorgan AIRSyntheticData team to generate 100K+ synthetic transactions using GANs + AutoNAS, achieving 96% accuracy and 0.97 AUC-PR for fraud detection while mitigating class imbalance by 40%	
NLP Architecture Analysis Parameter-efficient fine-tuning for sequence models Published, 1st Author	
Azure-Based Churn Analytics Large-scale predictive modeling & ETL pipelines Published, 1st Author	