Aniket Srinivasan Ashok

✉ a3sriniv@uwaterloo.ca | linkedin.com/in/aniket-srinivasan-ashok | github.com/aniketsrinivasan

Work Experience

Faire May. 2026 — Aug. 2026
ML Platform Engineer Intern *Waterloo, ON, Canada*

- Incoming Summer 2026.

Viggle AI (a16z) | *Python, Go, TypeScript, React, Docker* Jan. 2026 — Apr. 2026
Software Engineer Intern *Toronto, ON, Canada*

- Designed a distributed ML inference engine processing **70,000+ videos/day** across **8× GPUs**, enabling real-time person tracking/segmentation in complex videos with **95%+ accuracy** (ID-matching) at **20fps @ 1080p**.
- Implemented dynamic batching (YOLO/CLIP), yielding 3× **throughput**. Adapted Meta SAM3 with custom cache initialization, batch-inference support, and weight sharing, reducing per-video latency and VRAM by **40%**.

CaseClock @ Rockland 🔗 | *Python, Azure, FastAPI, LangChain, SQL, Git* Jan. 2025 — Apr. 2025
Software Engineer Intern (Founding Team) *Victoria, BC, Canada*

- Built voice-first AI platform from scratch, engineering a voice transcription + multi-agent serverless inference engine with Azure AI and Functions auto-scaling, serving **1,000+ requests/min** with <**5s** response time.
- Improved system performance achieving **4×** faster inference through asynchronous agent execution and semantic caching, and improved accuracy from **75% to 90%** using history-aware RAG with MCP tool-calling.
- Developed data pipeline with SQL, Azure Blob Storage, and MongoDB Atlas with automated ETL processes and document processing + vectorization. Reduced storage costs **20**× using SQL auto-archiving.

Unleash Networks 🔗 | *PyTorch, SQL, Git* May 2024 — Aug. 2024
ML Engineer Intern *Chennai, TN, India*

- Engineered time-series forecasting for network metrics achieving **sub-100ms** inference latency through ensembled XGBoost, LSTM, and ARIMA models, deployed on local servers using automated batched processing.
- Achieved >**94% accuracy** and <**1% false positive** on real-time unsupervised anomaly (DoS/DDoS) detection in multi-variate network data with variational autoencoders (VAEs) and Kolmogorov-Smirnov testing.
- Developed data cleaning and engineering pipelines for high-dimensional network data using Pandas and SQL.

Bradbury Group | *PyTorch, Slurm, Git, AWS* May 2025 — Present
ML Research Engineer *New York City, NY, USA (Remote)*

- Developed block-injection support for FLUX.1 and LLaMA to explore architecture model improvements.
- Ran **30+ ablations** on ViT models and developed deterministic block-replacement algorithms (including statistical correction, dynamic weight scheduling, etc.) for module distillation, enabling **15% faster convergence**.

Vision and Image Processing Lab @ UWaterloo | *PyTorch, Slurm, Git* Sep. 2024 — Aug. 2025
ML Researcher (Part-Time) *Waterloo, ON, Canada*

- Built and trained generative diffusion models (**200M–1.5B parameters**) for the **4**×, **8**× and **16**× super-resolution of wind data, reducing costs against traditional simulation methods by **over 100**×.
- Developed mechanisms for flow field reconstruction, including implementing flow matching schedulers, diffusion transformers (DiTs), iterative refinement algorithms, beam-search, physics-based and wavelet-domain losses.

Publications

1 **DCR: Fast and Stable Module Replacement in Transformers** | [OpenReview]
2 **Diffusion Models for Efficient and Accurate Super-Resolution of Wind Dynamics** | [arχiv]

Education

B. Math Data Science + Pure Math @ **University of Waterloo** Aug. 2023 — Apr. 2028

- **Coursework:** Algorithms, Data Structures, Stochastic Processes, Probability, NLP (Audit), Compilers

Technical Skills

Languages/Frameworks: Python; C/C++; ML Frameworks (PyTorch, WandB, HuggingFace, sk-learn, NumPy, Triton/CUDA); FastAPI, Jupyter, Colab, Slurm, Azure/AWS/GCP, MongoDB; SQL; Assembly (MIPS); SQL