

PROFILE

Perception and vision systems engineer developing AI tools/models locally: on-device, real-time, close to the data it acts on. Proven capability of driving products from prototype to live client operations. Deep focus on delivering impact aligned with key business priorities, leveraging modern vision systems, and prioritizing edge-capability and privacy.

SELECTED ACHIEVEMENTS

Research: ReVEAL (provisional patent, USPTO 2026, sole inventor): an auxiliary conditioning framework mitigating visual-attention decay in reasoning vision-language models, a lightweight attention-modulation approach rooted in time-series econometrics, adding under 550K parameters at constant memory and sequence-length cost. Third-party validation in progress; arXiv preprint in preparation. In pilot on animal-biomechanics video analysis.

EXPERIENCE

Lead AI Engineer

QRyde by HBSS, Lowell MA

Jan 2025 – Present

Transportation On-Device edge-hardware & Simulation:

- Led development of Q-DRIVE Safety Simulator + Edge Device, an AI/ML-enhanced driver-safety training and evaluation platform with an edge prototype custom-built Jetson NX in-vehicle device. Current beta deployment with a regional transit agency. Presented at three industry conferences (CTAA, APTA, CTA).
- Q-DRIVE Performance: 75% reduction in potential safety-related events per hour after 30+ sessions. Enrolled 96 transit agency drivers, 408 hours simulated data; 700 safety events logged over 231 sessions.
- Real-time evaluation and predictive future risk estimation using multi-tier scoring system extending NHTSA safety evaluation standards and ML pipeline for predictive future risk profiling; Safe-RL PPO policy in active development.
- Overhauled fragmented vision stack into a unified TensorRT CV pipeline, reducing end-to-end perception pipeline latency 7.5x (from 300ms to under 40ms at the 99th percentile) with sensor fusion on NVIDIA edge hardware. Deployed and in beta testing.

Edge Hardware CV & Cloud Integration:

- Launched a privacy-preserving edge AI facial recognition system with self-healing, reconnect & hot-load syncing. 30ms median detection latency, 14 recognized simultaneous faces @ 10 FPS. 99.9% uptime over 150 days.
- Stood up AWS GPU layer (FastAPI, Docker, vLLM) endpoint in a multi-function CV & AI schema, serving high-bandwidth vision-language models and ArcFace vectorizer.
- Collaborated with the IoT team to create a unified IoT workflow with edge-processing, AWS embedding vectorizer & SaaS WFM integration for automated attendance.
- Built a scalable, high-volume call-summarization (400+ calls per day) pipeline with multi-stage, programmatic and “LLM-as-Judge” validation pipeline.

Reinforcement Learning:

- Designing CMDP, MCTS dispatch optimization framework; running PUCT optimization trial.

Property Development Manager

J&J Asset Management, Boston, MA

Jan 2019 – Sep 2024

- Market selection and investment modeling for commercial, industrial, and multifamily development in emerging markets; vendor, contractor, and project management.

Statistician & Data Analyst

Ocean Associates, San Diego, CA

Sep 2016 – Aug 2018

- Contracted by NOAA to build time series ([S]ARIMA, geospatial & GLM models) for Pacific Bluefin tuna stock management, driving federal policy and annual catch quotas.

ENGINEERING PROJECTS

Multimodal Corpus Generation Suite · Data Pipeline

github.com/aiaxmaior/trainer_corpus_suite

- Modular Multimodal Analysis for VLM Corpus: 8+ concurrent stage processing pipeline featuring automated media preprocessing, CV, text analysis with 2-phase LLM-as-Judge validation.
- Systematic failure analysis of VLM captioning errors in this pipeline defined the problem space that spawned REVEAL.

Multimodal LoRA Training Harness · VLM Fine-Tuning

github.com/aiaxmaior/lora-harness-showcase

- Custom VLM fine-tuning harness (PEFT, HF Trainer) with explicit target-module selection across language and vision modules, assistant-only label masking, and cut-cross-entropy memory optimization; documents asymmetric multi-GPU tradeoffs. Built as the matched control for an attention-observation study.

AlphaZero-Style RL Engine · Reinforcement Learning

github.com/aiaxmaior/AlphaZero-lite

- Built an AlphaZero-style RL engine from scratch (PyTorch, MCTS); measured +300 Elo self-improvement, 20-0 area sweep of parent checkpoint in chess and applied it to simulated paratransit dispatch: on-time performance +28 pts, late minutes cut 9×.

SKILLS

Languages	Python (primary), C (production DeepStream, NVIDIA Jetson), SQL / PL-SQL
Edge & On-Device	NVIDIA Jetson (Orin Nano / NX), Jetson Platform Services, TensorRT, ONNX, DeepStream SDK (C/Python), CUDA, GStreamer, quantization (INT8 / NF4), inference optimization, latency budgeting
CV / Perception	Ultralytics/YOLO/SAM, ArcFace, MediaPipe, LaneATT, RTSP/IR camera pipelines, multi-object tracking, monocular distance estimation, sensor fusion (camera, IMU, GPS, ultrasonic)
AI / ML	PyTorch, Transformers, LoRA / QLoRA (PEFT), model serving (vLLM, TensorRT, ONNX), VLM / multimodal pipelines (Qwen, Gemma, Llama, ViT / CLIP / SigLIP), embeddings and vector-similarity retrieval, LLM judges, evaluation and failure-mode analysis
Services & Cloud	AWS, Docker, FastAPI, Pydantic, REST, WebSocket, IoT messaging, Linux, systemd, runtime monitoring
Quantitative	scikit-learn, pandas, GeoPandas, NumPy, GBM, time series (ARIMA / SARIMA, VAR), GLM, hypothesis testing / A/B evaluation, Oracle SQL, Spark / PySpark, R