Long Do

+1 (515) 947-6587 | workingliamdo@gmail.com | github.com/livecode111/

EDUCATION

University of South Florida Expected Graduation: **May 2026**

Bachelor of Science in Computer Science, Minor in Mathematics *Tampa, FL*

- **Coursework:** Algorithms, Computer Architecture, Operating System, Compilers, Deep Learning, Computer Vision, Hardware Accelerators for ML, CUDA Programming, Distributed System, Big Data Analytics, High Performance Computing, Embedded Systems, Introduction to Linux

EXPERIENCE

Amazon Robotics – *Incoming Software Development Intern* *Boston, MA* | **Aug 2025 – Dec 2025**

Coinbase – *Software Engineer Intern* *San Francisco, CA* | **May 2025 – Aug 2025**

- Built sandbox infra for Stablecoin B2B Payment API using Go, Docker, Kubernetes, for integration testing from partner wallets, merchants
- Developed mock server environment on AWS (EC2 + EKS) with Helm, ArgoCD, Datadog, simulating HTTP/gRPC responses, fault scenarios
- Routed external traffic through Envoy-based reverse proxies in entry gateway, MongoDB, securing services communication, observability

Scale AI – *Generative AI Intern* *Remote* | **Mar 2025 – May 2025**

- Built personalized ranking models using PyTorch, XGBoost to score 9M code interactions, improving most relevant retrieval accuracy by 18%
- Built retrieval-augmented generation pipeline in PyTorch, ElasticSearch vector stores to reduce LLM hallucination on 10M document queries
- Fine-tuned large language models (CodeT5, GPT) with mixed-precision training on DeepSpeed, improving code generation accuracy by 15%
- Deployed TorchX pipelines on AWS AI Platform for data validation, continuous training, Canary rollout, reducing model drift in production

Datacurve (YC W24) – *Machine Learning Engineer Intern* *Remote* | **Dec 2024 – Mar 2025**

- Trained distributed data pipeline using Python, Apache Spark, NVIDIA DALI on Linux, training code-repair models on 1K code repos
- Built transformer-based bug repair model using DeepSpeed ZeRO-3, WebDataset on 4 RTX GPUs with NCCL and gradient accumulation
- Implemented neural architecture search with JAX, Ray Tune for hyperparameter optimization, improving convergence by 40% on code corpora
- Created low-level CUDA, C++ kernels for FlashAttention inference, deploying models via VLLM, benchmarking performance via MLflow

University of South Florida – *AI Researcher* *Tampa, FL* | **Apr 2024 – Sep 2024**

- Developed hardware-aware neural architecture search using AutoTVM, TensorRT, for 5x inference speedup on transformer malware classifiers
- Built CUDA Graph reinforcement learning pipeline with cuDNN for threat detection, cutting training latency with kernel fusion, memory reuse
- Applied quantization-aware training, deployed ONNX export via Triton Server with model ensemble, dynamic batching, gRPC edge inference
- Benchmarked models on 15K malware samples using TensorBoard, Nsight System, and A/B testing on Linux, deploying via AWS Lambda

University of South Florida – *Computer Vision Researcher* *Tampa, FL* | **Aug 2023 – Dec 2023**

- Built self-supervised learning using contrastive loss (SimCLR), Pytorch to learn scene embeddings, boosting action recognition accuracy 15%
- Finetuned DeepLabV3+, MiDaS models for real-time semantic segmentation, depth estimation, improving 3D localization in urban scenarios
- Performed trajectory prediction with TensorFlow, Transformer spatiotemporal networks, achieving 92% accuracy on nuScenes motion forecast
- Trained imitation learning policy using expert demonstrations in CARLA, reducing trajectory deviation by 40% on held-out driving scenarios

PROJECTS

Robotic Perception & Behavior – *C++, OpenCV, PCL, ORB-SLAM3, PyTorch, Detectron2, ROS, DDPG, Perception*

- Built perception pipeline in C++, OpenCV, PCL, processing 8-camera + IMU streams for 6 DoF pose estimation with sub-cm drift over 500m
- Developed SLAM system using ORB-SLAM3 with GPU-accelerated bundle adjustment, cutting loop-closure detection latency by 70%
- Trained mask R-CNN model in Detectron2 on custom warehouse datasets, achieving 92% instance segmentation accuracy for bin-picking tasks
- Integrated reinforcement-learning grasping in PyTorch (DDPG algorithm) on NVIDIA Jetson AGX, improving pick success rate by 35%

High Performance Storage – *C++, SPDK, RDMA, eBPF, Protocol Buffers, NVMe*

- Designed storage engine in C++ using SPDK, implementing zero-copy I/O path and NVMe optimizations, reducing read/write latency by 70%
- Developed distributed cache with RDMA, memory allocator, lock-free data structures, scaling to handle 10K+ concurrent requests efficiently
- Built I/O monitoring system via eBPF (C++ bindings), kernel probes for realtime stack analysis, reducing bottleneck detection time by 60%

Quantum Edge High-Frequency Trading – *C++, RSI, SIMD*

- Developed market making system with lock-free order book and custom volatility estimation, maintaining 95% quote uptime with $15K+ PnL
- Engineered real-time risk management using SIMD-optimized Greeks calculations and position limiting, processing 100K+ updates/s
- Built adaptive pricing engine with technical indicators (RSI, SMA, EMA), inventory management, achieving 92% fill rate with <50µs latency

MarketMind – *Python, PostgreSQL, Redis, Kafka, Docker*

- Built real-time sentiment engine using scikit-learn and NLTK/TextBlob for ticker extraction, processing 100+ articles/second at 92% accuracy
- Developed high-frequency trading pipeline using Python/Apache Kafka for market data streams and sentiment correlation with <10ms latency
- Built ML pipeline combining market technicals (RSI, MACD) with BERT/DistilBERT ensembles, Docker, achieving 88% directional accuracy

TECHNICAL SKILLS

- **Tech:** C/C++, Python, Java, CUDA, PyTorch, TensorFlow, JAX, Ray | **Awards:** Meta Hacker Cup Round 2 (Top 1%), AMC Gold Medal