WESLEY JIN

AI Engineer | Senior Full-Stack Developer

U.S. Citizen | Fullerton, CA | wesley.busiiness@gmail.com | (415) 800-2802

Linkedin: linkedin.com/in/wesley-us | Github: github.com/eBlackrose | Portfolio: wesley-jin.vercel.app

SUMMARY

AI/ML Engineer and Senior Full-Stack Developer with 10+ years of experience architecting large-scale intelligent systems and production-grade AI platforms. Expert in multimodal and generative AI, on-device optimization, and distributed backend systems that power real-time, high-impact products. Proven record delivering AI-powered consumer experiences to 10M+ users at Nike, Google, and Uber. Passionate about bridging deep-learning research and scalable engineering to deploy state-of-the-art models across web, mobile, and embedded devices.

TECHNICAL SKILLS

AI & Machine Learning: PyTorch, TensorFlow, TensorFlow Lite, Core ML, TensorRT, Mediapipe, Computer Vision, NLP, LLMs, Multimodal AI, Generative Models, Direct Preference Optimization (DPO)
Programming Languages: Python, TypeScript, JavaScript, C++, CUDA, SQL
Frameworks & Libraries: React, Next.js, React Native, Node.js, Express.js, GraphQL
Cloud & DevOps: AWS (Kinesis, S3, Lambda), GCP (AI Platform), Docker, Kubernetes, CI/CD (GitHub Actions, Firebase)
Databases: PostgreSQL, MongoDB, DynamoDB, Redis
Hardware & Optimization Tools: NVIDIA Jetson, CUDA, TensorRT, OpenCV, CPack, On-Device Inference
Core Competencies: Multimodal AI, Edge AI, Generative AI, Model Optimization, Real-Time Inference, Distributed Systems, AI Infrastructure, Cloud Deployment

PROFESSIONAL EXPERIENCE

Senior AI & Full-Stack Engineer

Nike | Los Angeles, CA | *Mar 2023 – Present*

- **Architected and deployed** a multimodal AI motion analytics system powering Nike Run Club & Training Club, enabling **real-time performance feedback** for millions of global users.
- Designed full **edge-to-cloud pipeline**: on-device inference (TensorRT + NVIDIA Jetson) to data APIs (Node.js + GraphQL) to mobile & web visualization (React Native / Next.js).
- Developed custom **pose-estimation models (HRNet, Mediapipe)** with **18% higher precision** and **35% lower latency** via model quantization and distillation.
- Built **cross-platform mobile SDK** for live AI motion feedback using **TensorFlow Lite** and **Core ML**, integrated in iOS & Android apps.
- Integrated **AWS Kinesis + S3 + Lambda** for real-time streaming of thousands of frames/sec with autoscaling microservices.
- Mentored 5 engineers, standardized **TypeScript + GraphQL** design patterns across digital teams.
  Tech: Python, PyTorch, TensorFlow Lite, Core ML, Node.js, GraphQL, React Native, Next.js, AWS, TensorRT, NVIDIA Jetson

Full-Stack Web & Mobile Engineer

Postmates by Uber | Los Angeles, CA | *Jan 2020 – Feb 2023*

- Designed and built **Next.js + React Native** applications supporting **10M+ monthly active users**.
- Integrated **AI-driven ETA prediction** and **real-time courier tracking** using TensorFlow Lite and Google Maps API.
- Built **merchant analytics dashboards** (GraphQL + Node.js + PostgreSQL), improving query latency by **30%**.
- Migrated critical systems to **Uber's microservice platform** post-acquisition, increasing reliability and scalability.
- Automated CI/CD pipelines with **GitHub Actions + Docker + Firebase**, enabling weekly mobile releases.
- Collaborated with cross-functional design teams to enhance UX flows and user retention.
  Tech: React Native, Next.js, TensorFlow Lite, GraphQL, Node.js, PostgreSQL, Docker, Firebase

Software Engineer

Google | Sunnyvale, CA | *Jul 2018 – Dec 2019*

- Developed **AI-driven personalization systems** for **Google Assistant** and **Search**, improving contextual response accuracy.
- Built distributed **microservices for inference and data ingestion** handling millions of daily requests.
- Integrated **LLM-based NLP pipelines** using TensorFlow Extended (TFX) and GCP AI Platform.
- Refactored **REST APIs to GraphQL edge services**, reducing response latency by **25%**.
- Partnered with research teams to deploy **multimodal models (text + image)** into production.
- Led migration to **Kubernetes-based ML serving** with CI/CD rollout for model updates.
  Tech: Python, TensorFlow Extended, GCP AI Platform, GraphQL, Kubernetes, LLMs, Multimodal AI

AI Engineer Intern

Google | Berkeley, CA | *May 2017 – Jun 2018*

- Conducted **computer-vision research** to enhance Pixel mobile UI defect detection.
- Built hybrid **CNN-Transformer** models for temporal anomaly detection.
- Expanded datasets **10×** via automated augmentation pipelines.
- Reduced training time **from 1000 to 50 epochs** through optimized initialization.
- Co-authored internal research later integrated into **Pixel QA automation framework**.
  Tech: Python, Computer Vision, CNN, Transformer, TensorFlow, CUDA

AI Specialist Engineer

Pixellot (AI Sports Video Automation) | Palo Alto, CA | *Jun 2016 – Aug 2016*

- Reduced fisheye stitching latency for **4K video streams from 6s to 0.2s** via CUDA optimization.
- Optimized CUDA + OpenCV integration reducing processing from **0.2fps to 0.01fps**.
- Implemented **real-time 3D-to-2D video projection homographies** for player tracking.
- Delivered **6× faster OpenCV CUDA/CuDNN compilation** through CPack binary optimization.
- Built 360deg panoramic field reconstruction pipeline with real-time analysis (zoom, pan, slow motion).
  Tech: Python, CUDA, OpenCV, NVIDIA Jetson Xavier, Real-time Computer Vision

PROJECTS

Multi-Modal Generative AI Research

- Developed Direct Preference Optimization (DPO) achieving +9% performance over baseline models.
- Built self-supervised preference generation system removing need for human annotation.
- Reached 97% F1 on student engagement dataset using Visual QA optimization.
- Experimented with MiniGPT-4-Video and LLaVA-Next on DAiSEE, EngageNet, and SED datasets.
  Tech: Python, PyTorch, Computer Vision, NLP, DPO Algorithms

AI-Automated Sports Video Analytics

- Engineered real-time video analytics pipeline on NVIDIA Jetson Xavier.
- Reduced frame stitching time from 6s to 0.2s and CUDA integration from 0.2fps to 0.01fps.
- Built 360deg camera calibration and tracking system for automated sports coverage.
  Tech: Python, CUDA, OpenCV, NVIDIA Jetson

EDUCATION

University of California, Berkeley | College of Engineering
Bachelor of Science, Computer Science | *Aug 2014 – May 2018*

CERTIFICATIONS

AWS Certified Solutions Architect – Associate
AWS Certified Cloud Practitioner