Vladislav Kondratyev

vxk230059@utdallas.edu | LinkedIn: in/vladislav-kondratyev/ | GitHub: ch1kim0n1 | Certifications

EDUCATION

The University of Texas at Dallas | B.S. Computer Science, Minor in Business Administration | GPA: 3.92 | May 2027

- **Undergraduate Researcher:**
  - VR Behavioral Recognition & Brain Science Lab (**Dr. Ravi**) | Early Cancer Cell Detection Lab using AI (**Dr. Fang**)
  - Founder & Research Lead, EchoLab - authored research on AI memory systems (SECT model) for improved contextual reasoning
- Won 7 hackathons out of 20+, building AI-driven systems and full-stack applications under competitive constraints

WORK EXPERIENCE

Meta | *VR & AI Engineer* | *Austin, TX | Jan 2026 - Apr 2026 (Seasonal Contractor)*

- Architected and deployed a **real-time AI-powered VR system within Meta Horizon OS** integrating **PyTorch-based models** with immersive environments, enabling **context-aware agent interactions** through **low-latency inference pipelines (<100ms response time via A/B testing)**
- Designed and optimized a **distributed backend architecture (FastAPI, WebSockets, async processing)** to **handle real-time bidirectional data flow** between **VR clients and AI services**, supporting **scalable, concurrent user interactions**
- Developed **multimodal AI pipelines fusing user input, environmental context, and behavioral telemetry** into a shared embedding space via transformer-based encoders, driving adaptive agent behavior **with <100ms state update latency across concurrent VR sessions**

Intel Corporation | *Software Engineer Intern - XeSS Team* | *Austin, TX | May 2025 - Aug 2025*

- Exported and optimized **PyTorch super-resolution models to ONNX and TensorRT**, enabling cross-platform deployment; **applied INT8 quantization reducing model size by 40% with <2% quality degradation**
- Developed and optimized **deep learning–driven super-resolution models using PyTorch, OpenVINO, HLSL, and DPC++,** reducing temporal artifacts and **improving image stability by 25% in motion-intensive scenarios**
- Refactored and accelerated model inference pipelines **(C++, PyTorch, OpenVINO, CUDA)**, **reducing latency by 35%** and enabling real-time AI-driven **upscaling on Intel Arc GPUs**

Revent (Acquired, undisclosed) | *Founding Software Engineer* | *Bronxville, NY | Sep 2024 - Apr 2025*

- Co-Founded an **early-stage fintech startup (acquired)**, and built **a real-time financial processing system handling $300K+/month in transactions**, leveraging event-driven architecture for low-latency validation and processing
- Engineered **WebSocket-based streaming pipelines enabling live transaction monitoring** and **anomaly detection**, reducing discrepancies by 25%
- Implemented scalable **infrastructure and CI/CD pipelines (Docker, GitHub Actions), achieving 99.9% system availability** under high-frequency workloads

CNF Technologies | *Full Stack Engineer* | *San Antonio, TX | Aug 2024 - Mar 2026*

- Designed and optimized **distributed real-time decision pipelines using async event-driven architecture (Python, Java, Kafka)** improving throughput by **20% and reducing p99 end-to-end latency by 35% across high-frequency streaming workloads**
- Engineered **low-latency data transmission layers with zero-copy buffer management and connection pool tuning** enabling efficient handling of large-scale streaming data across **distributed microservices serving 15,000+ concurrent users**
- Diagnosed and resolved **system-level bottlenecks in Linux-based production environments via profiling (perf, strace, flamegraphs)**, **reducing downtime by 25%** and improving pipeline reliability under sustained high-throughput load

PROJECTS

EchoMind | May 2024 - Present

- Architected a **heterogeneous multimodal inference pipeline (audio via WebRTC, vision via OpenCV, OS telemetry via syscall hooks)** unified through a shared embedding space with async batching and **INT8/FP16 quantization achieving <150ms end-to-end latency**
- Engineered a **stateful agent runtime with attention-weighted cross-modal sensor fusion** and **sliding-window episodic memory buffer** enabling context-aware **adaptive behavior** across continuous streams without full context recomputation
- Designed an **event-driven inference backend (async Python, ZeroMQ pub/sub, gRPC tool dispatch)** supporting **parallel stream processing**, **dynamic tool execution, and low-overhead system integration via shared memory IPC**

Maes | Feb 2026

- Architected a **real-time AI agent system for intent-based computer control via speech using a Whisper-based STT pipeline** feeding a **fine-tuned LLM command interpreter** with structured JSON output parsing and **sub-200ms execution latency**
- Designed **multi-step agent workflows in LangChain/LangGraph with dynamic tool selection**, conditional branching, and stateful session memory to translate natural language intent into executable OS and browser actions
- Built a **cross-application automation layer (browser via Playwright, OS-level via pyautogui/Win32 API, REST APIs)** achieving **>95% task completion accuracy across 200+ real-world test scenarios**

EyeCore | Nov 2025 - Jan 2026

- Architected a **low-latency system telemetry platform in Rust using kernel-level APIs (procfs, /dev/input, inotify)** for **real-time collection and zero-copy streaming** of device-level signals to **downstream AI inference pipelines**
- Designed **high-throughput async data pipelines processing 10K+ system events/sec** with **lock-free ring buffers** and configurable sampling windows, sustaining <5ms capture-to-consumer latency under full load
- Exposed a **modular gRPC/REST API layer over collected telemetry streams enabling hot-pluggable integration** with external agents, adaptive systems, and real-time analytics consumers

TECHNICAL SKILLS

Languages: Python (primary), Rust (low-latency systems), C++, TypeScript, Go, SQL
AI & ML: PyTorch (training, fine-tuning, optimization), OpenVINO, TensorFlow, LightGBM, Transformers (HuggingFace, RoBERTa), CUDA kernel optimization, model evaluation, feature engineering, bias-variance tuning
Inference & Deployment: ONNX, TensorRT, OpenVINO, vLLM, real-time inference pipelines (<100ms), quantization (INT8/FP16), GPU workloads, scalable inference systems, SageMaker, batch inference strategies
Agentic & LLM Systems: LangChain, LangGraph, RAG pipelines, prompt engineering, embeddings, semantic search, vector databases (ChromaDB, FAISS, Qdrant), OpenAI/LLM APIs (tool calling, structured output parsing, multi-step workflows), function calling evaluation, agent evals & benchmarking, output validation & guardrails, memory architectures, token optimization, latency/cost tradeoffs
Real-Time Backend: FastAPI, WebSockets, async processing, gRPC, microservices, REST APIs, Redis (pub/sub, caching), agent-based systems, event-driven architecture
Data & Storage: PostgreSQL, MongoDB, vector storage, hybrid search (SQL + embeddings), DVC (data versioning)
Cloud & Infrastructure: AWS (EC2, Lambda, S3, SageMaker), GCP, Azure, Docker, Kubernetes, serverless deployment, GPU workloads
MLOps & Observability: CI/CD (GitHub Actions, Jenkins), Weights & Biases, MLflow, A/B testing, shadow deployments, model observability (drift/skew detection), pipeline orchestration, monitoring & performance tuning