

Zihan Dong

Machine Learning Researcher — Agents, LLMs, Multimodal ML, On-Device AI

Durham, NC | puma122707@gmail.com | (704) 951-5906

GitHub: [zdong104](#) | LinkedIn: [zihandong1004](#) | [Google Scholar](#) | [HomePage](#)

SUMMARY

Machine Learning Engineer specializing in agent systems built on LLMs, reinforcement learning, and multimodal memory. Focused on designing and deploying proactive, long-horizon agents capable of tool use, reasoning, and real-world task execution. Authored 10 research papers (NeurIPS / ICLR), filed 3 patents, and accumulated 200+ citations.

PROFESSIONAL EXPERIENCE

Lenovo – AI Ecosystem Engineer (Rotation Program)

Morrisville, NC | 2025 – 2026

- Built an end-to-end proactive agent system for long-horizon task execution with multimodal memory and tool execution for multi-step workflows, improving task success rate by 23% and reducing end-to-end latency by 5%
- Developed evaluation pipeline measuring task success rate, hallucination frequency, latency, and recovery behavior for long-horizon agent tasks
- Designed full agent loop (perception → memory → reasoning → tool execution → feedback) enabling long-horizon task completion and recovery
- Implemented tool-use / function-calling mechanisms enabling agents to interact with external systems and APIs

Lenovo – Workstation AI Solution Engineer

Morrisville, NC | 2024 – Present

- Improved agent-based automation workflows using LLM-driven task planning for system diagnostics and scheduling
- Implemented on-device LLM inference pipelines, optimizing GPU memory usage and reducing inference latency for workstation deployment
- Collaborated with architecture and systems teams to integrate agent runtime with workstation infrastructure
- Developed virtualization and hypervisor environments (e.g., Slurm, remote compute) for distributed AI workflows and model experimentation on Workstation

AINSIGHT – AI Sports Tracking Intern (CV Research & GPU Optimization)

Remote | Apr 2024 – Nov 2024

- Developed deep-learning-based sports tracking model using PyTorch; designed detection + pose + motion-tracking pipeline.
- Optimized CUDA and multi-GPU parallelization achieving 30%+ acceleration in inference latency.
- Built evaluation benchmarks and validated performance on diverse athlete datasets.
- Collaborated with the systems team to deploy the model to mobile clients and cloud GPU clusters (GCP/Azure).

North Carolina State University – Interactive Intelligence Research Intern

Remote | Apr 2024 – Oct 2024

- Built an AI assistant for manufacturing QA using Retrieval-Augmented Generation (RAG), combining text, images, tables, and figures.
- Conducted experiments in prompt engineering (CoT, ReAct) to improve reasoning and grounding across multiple LLMs.
- Identified failure modes in long-context and multimodal reasoning, and improved answer accuracy with a custom context database (DB Monge).

KPMG AI Lab – LLM & Vision Research Intern

Nanjing, China | May 2023 – Aug 2023

- Trained B2B computer vision models for document detection and segmentation, and fine-tuned LLMs for regulatory intelligence.
- Collected and cleaned large-scale legal documentation, creating a structured corpus for downstream CV and LLM tasks.
- Evaluated new SOTA models and recommended architectures to improve robustness and product quality.

PUBLICATIONS & PATENTS

Avoid Catastrophic Forgetting with Rank-1 Fisher from Diffusion Models — ICLR 2026

- Proposed a low-rank Fisher approximation to mitigate forgetting in diffusion model fine-tuning.
- Demonstrated significant reduction in performance decay under sequential domain adaptation.

ComputerAgent: Hierarchical RL Framework for General Computer Control — KDD 2026

- Designed hierarchical RL agent for GUI automation with multi-level action spaces and macro-actions, improving long-horizon task completion across applications
- Introduced structured macro-actions, screen-level state encoding, and cross-application generalization.

Additional publications: KDD, AAAI, ASA, JIMS (see Google Scholar)

Patent: ComputerAgent: A Hierarchical Reinforcement Learning Framework for General-Purpose Computer Control (IDF-175567)

Patent: Avoid Catastrophic Forgetting with Rank-1 Fisher from Diffusion Models (ID:63/888,843)

Patent: Self-Consolidating Language Models: Continual Knowledge Incorporation from Context. (ID:2026-488)

SELECT TECHNICAL PROJECTS

AI Sports Tracking (CV + CUDA)

- Built pose-tracking and movement segmentation model with multi-GPU CUDA parallelization.
- Delivered >30% latency reduction in inference for real-time sports analytics.

General-Purpose Agent for CAD & GUI Automation

- Built LLM + RL-based agent for CAD automation, generating action sequences for GUI-based design tasks and improving multi-step task completion

ZTrade: Trading Optimization App

- Built a mobile trading assistant using portfolio optimization and RL-inspired decision policies with real-time market signals.

Self-Replay: RL-based LLM Training with Limited Data

- Developed RL training pipeline for CAD **automation** and agent task execution.

TECHNICAL SKILLS

LLMs & Agents: tool use, function calling, ReAct, memory systems, long-context reasoning, agent evaluation

Multimodal: vision-language models, RAG, cross-modal retrieval

RL: hierarchical RL, policy learning, reward design

Systems: on-device inference, GPU optimization, distributed training (Slurm, Ray)

Languages: Python, JS/TS, Swift, Java (basic), SQL

Other: On-device inference, performance optimization, GPU debugging, embeddings, memory systems

EDUCATION

Georgia Institute of Technology – Atlanta, GA

Aug 2024 - May 2026

Master of Science in Computer Science (Specialization: ML / AI / Cognitive Systems)

GPA: 3.7/4.0

North Carolina State University – Raleigh, NC

Aug 2020 - May 2024

Bachelor of Science in Mechanical Engineering, Minor in Business Administration

GPA: 3.8/4.0

- NSF-funded researcher | Grand Challenge Scholar (Vice President)

LEADERSHIP & COMMUNITY

- Grand Challenge Scholar Program (President), NSF-funded researcher
- Member: I-Crops Entrepreneurship, Supply Chain Club, Business and Marketing Clubs
- NC State Chinese Student Association, App Development Club, Embedded ML Club
- University Men's Volleyball team

ADDITIONAL

- Languages: English (professional), Chinese (native), Japanese (basic)
- Interests: Volleyball, soccer, tennis, travel, diving, movies