

## Education

<b>Case Western Reserve University</b> , Cleveland, OH	Aug. 2024 – Present
Master of Science in Computer Science. Current GPA: 4.0/4.0; Thesis direction: Embodied AI, VLA, VLN. Coursework including: Machine Learning, Data Mining, Computer Vision, Machine Learning on Graph, Deep Generative Model.	
<b>University of Washington</b> , Seattle, WA	Sep. 2015 – Mar. 2020
Bachelor of Science in Applied Physics	

## Working Experience

<b>VULab</b> CWRU, Cleveland	May 2025 – Present
Research Assistant	
<ul style="list-style-type: none"><li>As the most junior level member on the team, presented research findings on Vision-Language-Action (VLA) models at <b>10+</b> internal research meetings, leading to new insights and identified <b>4+</b> potential areas for future research.</li></ul>	
<b>ZHIPU.AI (Z.ai)</b> , Beijing	May 2021 – Aug. 2024
Algorithm engineer, Part-time	
<ul style="list-style-type: none"><li>Developed models that enable metric-based document discovery and semantics-enhanced retrieval methods. These models were validated on <b>millions</b> of data entries and developed using Python, NumPy, the Neo4j Graph database, and MongoDB.</li><li>Implemented plugins and <b>Retrieval-Augmented Generation (RAG)</b> pipelines with the company's Large Language Model (GLM) for <b>3+</b> B2B solutions, including <b>LLM-enhanced retrieval</b>, <b>text-to-video</b>, and <b>image-to-video generation</b>.</li><li>Built <b>large-scale knowledge graphs</b> from journal articles and patents, comprising <b>over 90 million nodes</b> with an average degree of 12, to support metric computation and analysis.</li><li>Nominated for <b>7+</b> patents focused on big data classification and identification, NLP, and GNN-related methods.</li><li>Built and refactored a web application backend with optimized methods for executing asynchronous calculation tasks.</li><li>Managed the <b>terabyte-scale</b> databases, including <b>MongoDB to Elasticsearch</b> migration and database migration to Alibaba Cloud. Deployed and maintained local cluster CPU/GPU/containers monitoring using <b>Grafana</b>.</li><li>Implemented <b>in-memory</b> relational storage using <b>Redis</b> bitmaps and data stream processing using <b>RabbitMQ</b>. Designed and proposed a microservices architecture, boosting online system performance by <b>7x</b>.</li><li><b>Led and mentored</b> a team of 4 interns, driving consistent performance and fostering strong cross-functional collaboration with the data department.</li></ul>	
<b>Founder Securities Co., Ltd.</b> , Beijing	Sep. 2020 – Dec. 2020
Quantitative analyst intern	
<ul style="list-style-type: none"><li>Developed quantitative trading strategies in Python, including a multi-factor model based on research reports and a statistical model targeting on northbound Hong Kong capital flows affecting the mainland A-share market.</li></ul>	

## Personal Project Experience

<b>An Agentic Navigation Framework Utilizing Vision-Language Models</b> , Cleveland	Jan. 2025 – May 2025
<ul style="list-style-type: none"><li>Designed and implemented a Vision-Language-based <b>navigation agent</b> leveraging Qwen 2.5-VL (7B) as a cognitive “main brain” with structured memory and reflective reasoning. Engineered custom system + user prompting strategies to enable contextual planning and semantic understanding for embodied tasks. Integrated and evaluated the framework within Habitat-Lab and Room-to-Room (R2R) environments, demonstrating improved instruction following and environment grounding. <a href="#">Arxiv 2506.10172</a></li></ul>	
<b>Enhancing Video Retravel Using VLM</b> , Cleveland	Oct. 2024 – Dec. 2024
<ul style="list-style-type: none"><li>Designed and developed a scalable backend and database layer for an application focused on retrieving relevant videos based on video or image input. Responsibilities include engineering on VLM model Qwen 2-VL (7b), retrieval method, and Vector DB integration. Used Transformer, Neo4j, and Pinecone, etc. for implementation. <a href="#">Arxiv 2503.17415</a></li></ul>	
<b>Low-Rank Adaptation Defense with Robustness</b> , Cleveland	Oct. 2024 – Dec. 2024
<ul style="list-style-type: none"><li>Developed and evaluated a LoRA-based defense pipeline for a ResNet-18 model to counter Feature Importance Attacks (FIA). Achieved a best validation adversarial accuracy of 98.60% within only 2 epochs, demonstrating rapid model robustness improvement. Used Pytorch for implementation.</li></ul>	

## Skills

**Programming Languages:** Python, Linux Bash. **ML Frameworks:** Transformer, Pytorch, Numpy, Scikit-learn, Triton **Database Management Systems:** MongoDB, Elasticsearch, Kibana, Neo4j, Pinecone, PostgreSQL, Redis. **Message Queues:** RabbitMQ, Kafka. **Container:** Docker, Grafana. **General:** API/REST, Flask, Django, SQL, GitHub, Agile, DevOps, Ansible, Profiling, AWS, CV, LLM, Product management, Leadership, SolidWorks, building drones.

## Accomplishments

<b>Patents</b>	May 2021 – Present
Document fining methods in large corpus: <a href="#">CN 114510584 B</a> <a href="#">CN 114969251 A</a> <a href="#">CN 115471483 A</a> ; Topic crusting: <a href="#">CN 116644338 B</a> <a href="#">CN 116561605 B</a> ; Information Retrieval: <a href="#">CN 117216417 B</a> ; Generating training patterns: <a href="#">CN 118277794 A</a> ;	