My pursuit of a Master's in Computer Science at the University of Alberta is driven by my academic and professional experiences, which have highlighted the critical need to deepen my understanding of ML fundamentals. As an undergraduate researcher at the Programmable Reality Lab at University of Calgary under Prof. Ryo Suzuki(now at the University of Colorado-Boulder), I developed an augmented reality system leveraging Google's MediaPipe for computer vision tasks. While this project demonstrated the power of ML as a tool, it also revealed my limited grasp of the underlying algorithms—a gap further underscored in my current role as an AI Infrastructure Engineer at Huawei's Ascend Computing division. In this role, I develop and optimize low-level software for AI accelerators but recognize that advancing next-generation AI systems requires expertise in core ML principles, particularly deep neural networks and reinforcement learning—areas pioneered by notable faculty within the Alberta Machine Intelligence Institute (Amii), such as Prof. Richard Sutton and Prof. Michael Bowling. Through this program, I aim to bridge my systems engineering background with advanced ML theory, equipping myself to lead research-driven development in industry.

In my role as an AI Infrastructure Engineer at Huawei's Ascend Computing division, I specialize in developing high-performance software for training and deploying start-of-the-art large language models on Huawei's Ascend NPU hardware. My work focuses on three key areas: (1) designing and optimizing communication primitives in HCCL (Huawei's Collective Communication Library) to compete with NVIDIA's NCCL, (2) developing compute kernels in Ascend C (Huawei's hardware-specific parallel computing language), and (3) contributing to Huawei's PyTorch-NPU adapter to enable efficient model execution. Through implementing fundamental operations like matrix multiplication and activation functions (ReLU, softmax) at the hardware level, I have gained intimate knowledge of transformer architectures and model quantization techniques.

However, this systems-level work has revealed a fundamental truth: hardware optimization alone cannot overcome algorithmic limitations. A striking example is the evolution from AlexNet to SqueezeNet, where SqueezeNet's architectural innovations reduced parameters by 50x without hardware changes, all while achieving similar ImageNet accuracy. In my own work optimizing Huawei's MINDIE inference engine, I encountered similar constraints: no amount of kernel optimization could compensate for inefficient model architectures. When deploying Alibaba's Qwen model, we achieved only marginal speedups (1.2x) through quantization, while algorithmic improvements in later versions delivered 3x gains with comparable accuracy. This pattern persists in distributed training via HCCL, where communication overhead often stems from model architecture choices rather than hardware limitations. These experiences have convinced me that next-generation AI systems require co-design of algorithms and infrastructure, a perspective I aim to develop through University of Alberta's Master of Science in Computer Science program.

While my undergraduate curriculum at the University of Calgary emphasized human-computer interaction over formal ML training, I proactively cultivated research skills that directly translate to machine learning. As first author of a UIST 2023 paper[1], I designed a real-time AR collaboration system that required iterative hypothesis testing—a methodology central to ML research. This involved statistical validation of user interactions and peer-reviewed critique of system performance, mirroring the evaluation rigor needed for ML model development. My honors thesis[3] further demonstrates this translational capacity: inspired by Prof. Joel Reardon's USENIX Security Best Paper[4], I reverse-engineered 35,117 Android applications to build a labeled dataset of 11,978 privileged API calls. This project demanded large-scale data engineering (creating pipelines for APK decompilation and call-graph analysis) and feature engineering (identifying security-relevant patterns)—skills directly applicable to ML data preprocessing. While my institution's curriculum lacked dedicated ML coursework, I developed its essential research competencies through applied projects—a capability proven by my first-author publication at the top-tier ACM UIST 2023 conference, where I successfully formulated a research question, designed a rigorous validation protocol, and managed a complex implementation dataset.

Crucially, both my undergraduate research experiences revealed the limitations of non-ML approaches. In the HCI project, if it were not the object tracking ability provided by Mediapipe, the system would not

achieve high expressiveness and made such innovation and contribution. For my honors thesis, I found that manual analysis of API privileges did not scale, as it required inspecting reverse-engineered bytecode from every application—even though most are benign. This process could be significantly improved and scaled by employing a machine learning model to predict maliciousness, reserving deep analysis only for high-risk cases. These gaps motivated my transition to ML: I have demonstrated my ability to execute rigorous research in computer science and made two publication RealityCanvas[1] and RealityEffect[2], and now seek the mathematical foundations and algorithmic tools (through University of Alberta's program) to solve such problems with machine intelligence.

My academic record also demonstrates strong preparation for graduate studies in machine learning, including a 3.92/4.0 GPA in Computer Science at the University of Calgary. Recognizing that advanced machine learning is fundamentally built upon linear algebra and multivariate calculus, I dedicated my elective credits to mastering these foundations, excelling in Calculus I-III, Linear Algebra, and Advanced Algebra. This mathematical foundation provided the formal reasoning skills essential for theoretical ML. For instance, my linear algebra knowledge was directly applied at Huawei to manipulate high-dimensional data through tensor operations, while calculus was crucial for understanding loss function optimization and gradient descent.

My connection to the University of Alberta began when I was offered admission as a transfer student after my first year. While this opportunity was interrupted by the COVID-19 pandemic, I now seek to join the Master of Science in Computer Science program, bringing significantly enhanced research experience, work experience and clear goals: to bridge my systems expertise from Huawei's AI infrastructure and HCI publications with deeper machine learning knowledge under faculty such as Prof. Sutton (reinforcement learning), Prof. White (efficient systems) and others. The program's balance of theoretical rigor and practical application perfectly supports my transition from implementing models to designing innovative ML models, as I aim to both master foundational algorithms through coursework and gain hands-on model design experience through thesis research targeting top-tier conferences (NeurIPS, ICLR, ICML). By combining my production-level development skills with University of Alberta's world-class ML curriculum, I will be positioned to contribute meaningful research while advancing from AI system optimization to algorithmic innovation.

[1] Xia, Zhijie, et al. "RealityCanvas: Augmented Reality Sketching for Embedded and Responsive Scribble Animation Effects." Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. 2023.

[2] Liao, Jian, et al. "RealityEffects: Augmenting 3D Volumetric Videos with Object-Centric Annotation and Dynamic Visual Effects." Proceedings of the 2024 ACM Designing Interactive Systems Conference. 2024.

[3] Xia, Zhijie, "Invesigate Android Permission System" Manuscript. 2023.

[4] Reardon, Joel, et al. "50 ways to leak your data: An exploration of apps' circumvention of the android permissions system." 28th USENIX security symposium (USENIX security 19). 2019.