Research Proposal: New Inference Paradigms for ML at the Edge

Jinming Ren

November 4, 2025

Abstract

This project targets real-time, on-device ML through RISC-V + FPGA co-design and neuromorphic inference. We first optimize an ANN detector (e.g., RT-DETR) via sub-8-bit quantization and CFU-based acceleration to achieve deterministic low latency. We then develop an SNN path with few time-steps, plus a portable SNN IR and quantization API for cross-backend deployment. We study internal complexity (richer neuron dynamics) to replace external depth/width and explore compute-in/near-memory for memory-bound kernels. Using NeuroBench-style evaluation on video and event-camera tasks, we aim for <10 ms latency and $\geq 30 \sim 50\%$ energy savings at $\leq 1 \sim 2\%$ accuracy drop versus a GPU-edge baseline, releasing all artifacts for reproducibility.

1 Motivation

Over the past decade, cloud-based training and inference pipelines face growing issues of **high latency, bandwidth bottlenecks, data privacy issues** [1] and escalating **training energy cost** [2], etc. Edge GPUs partially alleviate these issues but, on the one hand, remain too general-purpose, lacking flexibility for custom numeric precisions, memory hierarchy and data movement [3]. On the other hand, different applications stress different aspects:

- **Autonomous systems, UAVs, Controlled nuclear fusion**: Demand sub-millisecond latency for control stability and decision safety [4, 5, 6].
- **Medical and IoT devices**: Prioritize data privacy and local analytics to meet regulatory and ethical standards [1].

Recent evidence [7] shows **FPGA-based accelerators** can outperform GPUs in both **energy efficiency** and **deterministic latency** when tuned for application-specific workloads. Therefore, FPGA-based hardware-software co-design emerges as a promising solution for low-power, low-latency edge computing tailored to specific application needs.

2 Problem Statement & Research Questions

Problem. Edge AI is constrained by end-to-end latency, energy per inference, and privacy. General-purpose accelerators lack support for application-specific numerics and event-driven workloads. SNNs promise ultra-low-power inference but lack portable toolchains (IR/quantization) and competitive accuracy at low time-steps.

1. How far can RISC-V + FPGA co-design push latency/energy for real-time perception while preserving accuracy?
2. Can SNNs with internal complexity (multi-timescale, adaptive thresholds) match ANN accuracy at few-time-step inference on FPGA?
3. What portable SNN IR + quantization API enables "write once, target multiple backends" without accuracy regressions?
4. Which dataflows (streaming vs memory-centric) minimize data motion for attention/convolution bottlenecks at the edge?

3 Project Objectives & Expected Contributions

1. An open source, reproducible full-stack edge ML system (models → compiler → RTL → bitstream) that surpasses the ANN GPU-edge baseline on ≥ 2 of latency, energy per inference, accuracy for a selected application scenario.

2. A portable SNN IR + quantization API for spike dynamics, demonstrated on FPGA.

3. Evidence that internal complexity can reduce depth/width and time-steps while maintaining accuracy on edge tasks.

4. An open benchmark harness with NeuroBench-style reporting for fair cross-hardware comparison.

There are two variables to choose though:

- **Application scenario**: Suits edge computing and extremely low latency (possibly video stream processing / controlled nuclear fusion).

- **Target network structure**: Possibly accelerating RT-DETR [8] or YOLO, or directly optimize existing Spiking Neural Networks (SNNs).

4 Methodology

To achieve the project objectives, the proposed research will follow a three-stage methodology: The warm-up phase establishes an ANN baseline and a RISC-V-coupled FPGA accelerator with deterministic low-latency dataflows. The mid-term phase investigates brain-inspired computing, especially SNNs, focusing on improving accuracy, enhancing LIF model (increasing "internal complexity" [9, 10]), sparse computation, quantization / pruning models and establishing SNN standard APIs. The long-term phase explores emergent intelligence by investigating hypercomputation beyond Turing limits. The details are as shown in the following sections.

4.1 Warm-up: Von Neumann path (RISC-V + FPGA, ANN first)

This phase has been planned as my graduation project, with the aim to pushing the limits of von Neumann architecture (RISC-V) in edge computing by accelerating and optimizing existing ANNs. Take target network RT-DETR [8] as an example, the steps are as follows:

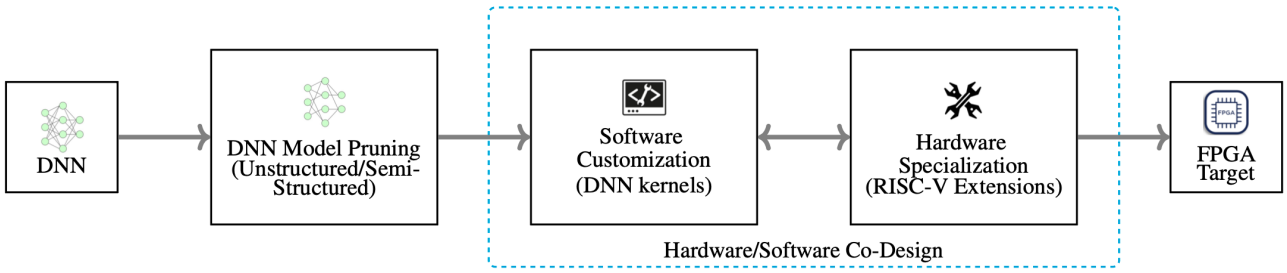

Figure 1: Warm-up phase design flow (adapted from [11])

- **Quantization & pruning for edge**: We adopt sub-8-bit quantization (INT8 → INT4/INT2 if accuracy allows) and structured sparsity to minimize off-chip transfers. We will report accuracy-latency-energy trade-offs under identical datasets and input resolutions.

- **RISC-V-based accelerator with CFU Playground**: As a undergraduate warm-up project, we will imitate the design flow in [11] as shown in Figure 1. We will implement a VexRiscv + Custom function unit (CFU) design within the open-source CFU Playground framework [12] shown in Figure 2. The difficulties are to design customized extended instructions in RISC-V for computational-intensive operators (e.g., depthwise/pointwise convolution, QK^T, low-bit GEMM), write the corresponding RTL for the hardware.

- **Full-stack design and DSE**: As a learning experience, we will reinvent the wheel by designing the entire accelerator in `Chisel` from scratch and open-sourcing it on Github. Then, we also explore a large multi-dimensional design space exploration (DSE) using automated methods (such as heuristic or evolutionary algorithms) to identify optimal configurations balancing accuracy, energy, and latency. Finally, we will use Arty A7-100T FPGA as the hardware platform for real measurements.

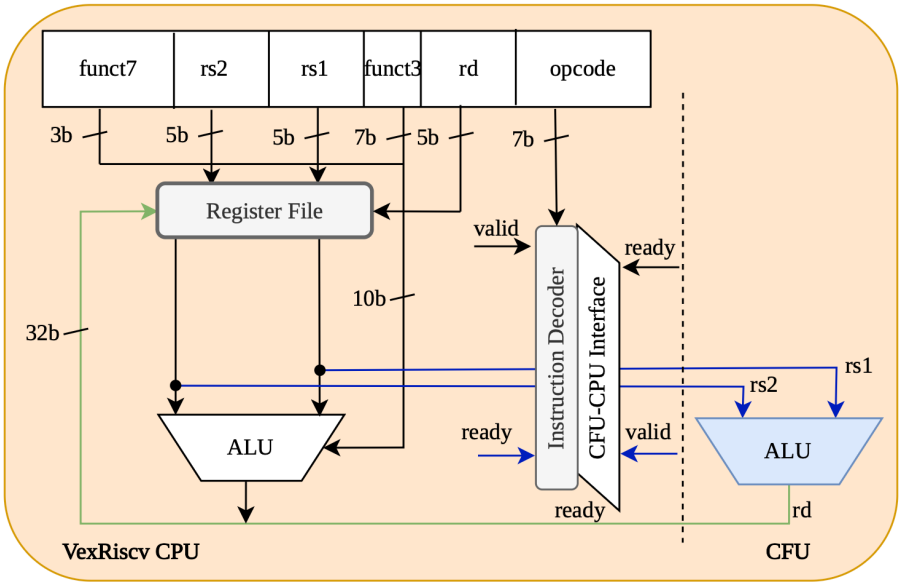

Figure 2: CFU-playground: ML Accelerators in RISC-V ISA [11, 12]

4.2 Mid-term: Brain-Inspired path (SNN on FPGA, portable toolchain)

This phase is planned to be my potential PhD research topic. We will explore Brain-Inspired Computing (BIC) in computer vision tasks with a particular focus on EdgeSNNs (parameters < 100 M [1]) and In-Memory Computing (CIM).

SNNs have shown promise for ultra-low-power, event-driven inference at the edge. SNNs model neurons with explicit membrane dynamics (LIF model as shown in Figure 3). Unlike conventional ANNs, SNNs process information in the temporal domain using binary spikes (event-driven coding), this is particularly suitable for SpikeCV cameras [13], which in my view are the next-generation vision sensors for edge applications such as autonomous driving.

Previous work on SNNs in autonomous driving includes Spiking-YOLO [14], EMS-YOLO [15], etc. However, the performance is still not good in general compared to ANN [16]. ECS-LIF in [16] suggests high-performance spiking detectors are possible with ANN-level matched accuracy and ultra-low energy costs. However, one still need to carefully select hardware architectures. For video classification, research is even more nascent [13]. Early attempts include spiking recurrent networks (processing up to 300 time steps of video frames), or hybrid ANN-SNN approaches for action recognition [13], SpikeVideoTransformer [17], SpikeYOLO [18]. However, there is no SNN equivalent yet for many popular video models (e.g. no published spiking variant of SlowFast or DETR detection-transformer as of 2025).

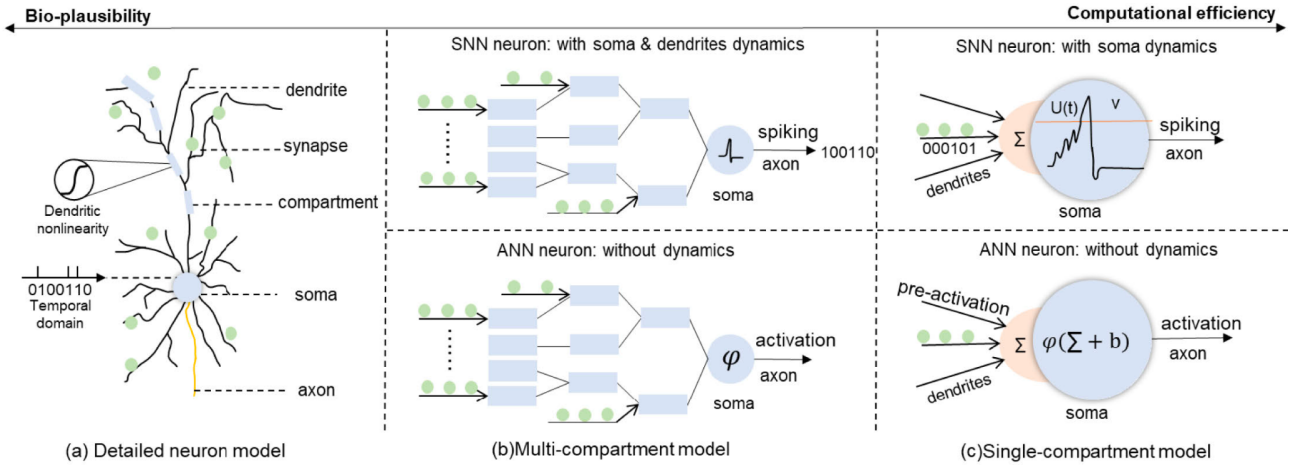

Figure 3: Leaky integrate-and-fire (LIF) neuron dynamics in SNN [10]

Since the research steps are not as clear as the warm-up phase, I list some potential research directions in parallel below:

- **Explore SNN design / training method**: There are two routes to get optimized SNNs: ANN2SNN conversion [8, 1], or direct-training SNNs. We will explore both routes to find the best-performing SNNs for our target application. We will also investigate advanced training techniques such as surrogate gradients, temporal backpropagation, and biologically inspired learning rules to improve SNN performance.
- **Develop unified SNN toolchain**: SNNs lack standard APIs and SNN-specific IRs (analogous to ONNX [1, 19]) for quantization strategies tailored to spike dynamics [19]. We prototype a minimal graph-level SNN IR (ops, neuron nodes, timing semantics) plus a quantization API (e.g., INT8 × INT2 spike ops, ternary spikes, per-layer time-step budgets) to decouple front-end training from back-end compilers, following the direction of NIR and recent co-design work that targets FPGA-friendly spike arithmetic. This addresses today's fragmentation across neuromorphic stacks and enables "write once, target FPGA/Loihi/MCU." [20]
- **Further explore SNN internal complexity**: A recent study [9, 10] by Network model with internal complexity bridges artificial intelligence and neuroscience shows a pivotal shift in thinking: Instead of simply growing neural networks by adding more layers or parameters ("external complexity"), we can embed richer dynamics *inside* each neuron or module — a paradigm the authors term "small model with internal complexity".

4.3 Long-term: Emergent Intelligence

This phase is my long-term aspiration toward artificial general intelligence (AGI). This stage will explore the theoretical and practical computational boundary and brand-new distributed computing paradigms inspired by the human brain under the guidance of recent theoretical investigations into *emergence* in artificial systems, such as Berti et al. (2025) [21] who survey emergent abilities in LLMs and identify conditions like scaling, criticality and compression that contribute to spontaneous capability gains. Continuing the exploration of internal complexity in Section 4.2, there are two major directions: Turing-complete machine and hypercomputation beyond Turing limits.

- **Decentralized, event-driven architectures**: We will first explore turing-complete hardware and software systems that mimic the *highly-distributed, asynchronous nature and learning-while-inferencing feature* of the brain. *Turing-equivalent* cellular automata

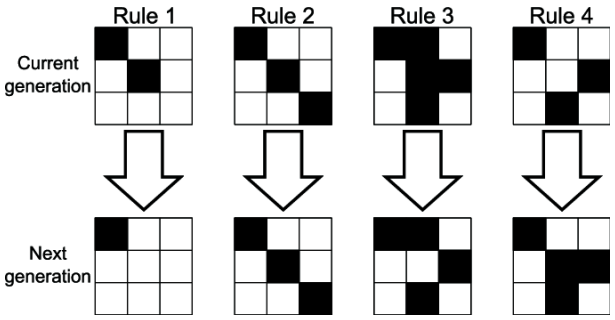

Figure 4: Conway's Game of Life (CGOL): local update rules [22].

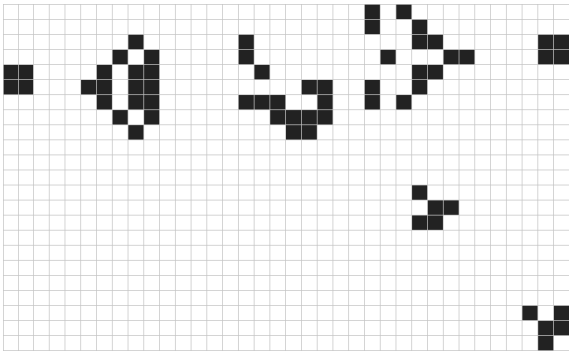

Figure 5: Gosper's glider gun: a self-replicating pattern proving Turing-completeness [23, 24].

such as CGOL [25] (Figure 4), Langton's ant, Particle Life already demonstrate how simple local rules can give rise to complex, emergent patterns (Figure 5). While using the CGOL itself as a practical "computer" is inefficient, it serves as a proof-of-concept that emergence can arise from simple components. The challenge is discovering the *right set of rules (i.e., internal complexity) or learning algorithms* that yield robust emergent intelligence, not just (external) complexity for its own sake [26]. Then turn out to nature (the hardware) to find out if there is a machine under our control that performs this set of rules intrinsically.

- **Hypercomputation beyond Turing limits**: Human brain might be exploiting computational principles beyond the scope of traditional Turing machines. Penrose and others (like Stuart Hameroff) have hypothesized that quantum effects in neural microstructures (e.g. microtubules) could enable the brain to do things standard computers cannot [27]. Achieving AI with brain-like cognition might then require tapping into quantum computing [28], ONNs [29], Organoid Intelligence (OI) [30, 31] and beyond.

References

[1] Shuiguang Deng, Di Yu, Changze Lv, Xin Du, Linshan Jiang, Xiaofan Zhao, Wentao Tong, Xiaoqing Zheng, Weijia Fang, Peng Zhao, Gang Pan, Schahram Dustdar, and Albert Y. Zomaya. Edge intelligence with spiking neural networks, 2025.

[2] Chapter 1: Research and development, 2025.

[3] Asking Questions. Introduction to fpgas and ml inference with hls4ml (benjamin ramhorst, 8 november 2024) - youtube. 2025.

[4] Safa Mohammed Sali, Mahmoud Meribout, and Ashiyana Abdul Majeed. Real time fpga based cnns for detection, classification, and tracking in autonomous systems: State of the art designs and optimizations, 2025.

[5] Javier Duarte et al. Fast inference of deep neural networks in FPGAs for particle physics. *JINST*, 13(07):P07027, 2018.

[6] Fusion Energy Sciences. Ai tackles disruptive tearing instability in fusion plasma | department of. 2025.

[7] Feng Yan, Andreas Koch, and Oliver Sinnen. A survey on fpga-based accelerator for ml models, 2024.

[8] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection, 2023.

[9] Linxuan He, Yunhui Xu, Weihua He, Yihan Lin, Yang Tian, Yujie Wu, Wenhui Wang, Ziyang Zhang, Junwei Han, Yonghong Tian, Bo Xu, and Guoqi Li. Network model with internal complexity bridges artificial intelligence and neuroscience. *Nature Computational Science*, 4(8):584–599, 2024.

[10] Guoqi Li, Lei Deng, Huajin Tang, Gang Pan, Yonghong Tian, Kaushik Roy, and Wolfgang Maass. Brain-inspired computing: A systematic survey and future trends. *Proceedings of the IEEE*, 112(6):544–584, 2024.

[11] Muhammad Sabih, Abrarul Karim, Jakob Wittmann, Frank Hannig, and Jürgen Teich. Hardware/software co-design of risc-v extensions for accelerating sparse dnns on fpgas, 2025.

[12] Shvetank Prakash, Tim Callahan, Joseph Bushagour, Colby Banbury, Alan V. Green, Pete Warden, Tim Ansell, and Vijay Janapa Reddi. Cfu playground: Full-stack open-source framework for tiny machine learning (tinyml) acceleration on fpgas. In *2023 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, page 157–167. IEEE, April 2023.

[13] Yasser Ashraf, Ahmed Sharshar, Velibor Bojkovic, and Bin Gu. Spact18: Spiking human action recognition benchmark dataset with complementary rgb and thermal modalities, 2025.

[14] Seijoon Kim, Seongsik Park, Byunggook Na, and Sungroh Yoon. Spiking-yolo: Spiking neural network for energy-efficient object detection, 2019.

[15] Qiaoyi Su, Yuhong Chou, Yifan Hu, Jianing Li, Shijie Mei, Ziyang Zhang, and Guoqi Li. Deep directly-trained spiking neural networks for object detection, 2023.

[16] Miao Jin, Xiaohong Wang, Ce Guo, and Shufan Yang. Research on target detection for autonomous driving based on ecs-spiking neural networks. *Scientific Reports*, 15(1):13725, 2025.

[17] Shihao Zou, Qingfeng Li, Wei Ji, Jingjing Li, Yongkui Yang, Guoqi Li, and Chao Dong. Spikevideoformer: An efficient spike-driven video transformer with hamming attention and $\mathcal{O}(t)$ complexity, 2025.

[18] Xinhao Luo, Man Yao, Yuhong Chou, Bo Xu, and Guoqi Li. Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection, 2025.

[19] Alessio Carpegna, Alessandro Savino, and Stefano Di Carlo. Spiker+: A framework for the generation of efficient spiking neural networks fpga accelerators for inference at the edge. *IEEE Transactions on Emerging Topics in Computing*, 13(3):784–798, July 2025.

[20] Jens E. Pedersen, Steven Abreu, Matthias Jobst, Gregor Lenz, Vittorio Fra, Felix Christian Bauer, Dylan Richard Muir, Peng Zhou, Bernhard Vogginger, Kade Heckel, Gianvito Urgese, Sadasivan Shankar, Terrence C. Stewart, Sadique Sheik, and Jason K. Eshraghian. Neuromorphic intermediate representation: A unified instruction set for interoperable brain-inspired computing. *Nature Communications*, 15(1):8122, 2024.

[21] Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci. Emergent abilities in large language models: A survey, 2025.

[22] Takayuki Hirose and Tetsuo Sawaragi. Extended fram model based on cellular automaton to clarify complexity of socio-technical systems and improve their safety. *Safety Science*, 123:104556, 03 2020.

[23] Ivan Lokhov. Game of life. 2025.

[24] Wikipedia Contributors. Gun (cellular automaton), 08 2023.

[25] James McCrum and Terence P Kee. Conways game of life as an analogue to a habitable world livingness beyond the biological, 2024.

[26] David C. Krakauer, John W. Krakauer, and Melanie Mitchell. Large language models and emergence: A complex systems perspective, 2025.

[27] Darren Orf. Does quantum entanglement in your brain generate consciousness? 2025.

[28] Gabriel A. Silva. Leveraging quantum superposition to infer the dynamic behavior of a spatial-temporal neural network signaling model, 2025.

[29] Tingzhao Fu, Jianfa Zhang, Run Sun, Yuyao Huang, Wei Xu, Sigang Yang, Zhihong Zhu, and Hongwei Chen. Optical neural networks: progress and challenges. *Light: Science & Applications*, 13(1):263, 2024.

[30] Lena Smirnova, Brian Caffo, David Gracias, Qi Huang, Itzy Erin Morales Pantoja, Bohao Tang, Cynthia Berlinicke, J. Boyd, Timothy Harris, Erik Johnson, Brett Kagan, Jeffrey Kahn, Alysson Muotri, Barton Paulhamus, Jens Schwamborn, Jesse Plotkin, Alexander Szalay, Joshua Vogelstein, and Thomas Hartung. Organoid intelligence (oi): the new frontier in biocomputing and intelligence-in-a-dish. *Frontiers in Science*, 1:1017235, 02 2023.

[31] Timothy Oh. The future of artificial intelligence — wetware. 2025.