

# Yuhao Ge

Mobile: (217)926-3291 | Email: [yuhao2@illinois.edu](mailto:yuhao2@illinois.edu) | [Homepage](#) | [LinkedIn](#)

## Education

<b>University of Illinois at Urbana-Champaign</b>	M.Sc. Computer Science	GPA: <b>4.0</b> /4.0	2023.8 - 2025.5
<b>University of Illinois at Urbana-Champaign</b>	B.Sc. Computer Engineering	GPA: <b>3.95</b> /4.0	2019.8 - 2023.5
■ Honors: Highest Honors, Bronze Tablet (3%, 2023), Dean's List (2020&2022)			
<b>Zhejiang University</b>	B.Eng. Computer Engineering	GPA: <b>3.93</b> /4.0	2019.8 - 2023.5
■ Honors: Zhejiang Provincial Government Scholarship, First-Class Scholarship (3%, 2020&2022)			

## Skills

- **Programming:** Python, C/C++, JavaScript, SystemVerilog, Assembly, SQL, MongoDB, Neo4j
- **Frameworks & Tools:** CUDA, PyTorch, TensorFlow, TVM, Triton, PyG, Flask, React, Node.js, Kubernetes, AWS, GCP

## Work Experience

- Amazon, Annapurna Labs** | *Software Engineer Intern* | *Compiler, Systems, ML, Accelerators* 2024.5 – 2024.8
- [Neuron Compiler](#) | *Optimize deep learning on AWS AI accelerators (Trainium and Inferentia)*
    - Delivered **three** organization-wide presentations and received a **Certificate of Appreciation** for developing one of the most influential internship projects, which significantly enhanced performance on Trainium and opened up a new direction for future development.
    - Developed infrastructure for automatic kernel generation, compilation, profiling, and visualization with a defined sweep space
    - Collected data for **DMA access pattern** analysis and introduced a **learning-based DMA latency model**
    - Created the **first-generation Autotuning** infrastructure from scratch for **compiler optimization** and **kernel optimization**
    - Used autotuning to optimize the Matrix Multiply **Fusion Pass**, achieving a 14.7% improvement for the **Llama3.1** model
    - Developed **kernel language** for AI accelerators and supported the **kernel optimization** with autotuning, achieving a 4.9% HFU improvement for kernels like Matrix Multiply
    - Implemented **multi-process compilation** and **distributed benchmarking**, resulting in 8.62X speedup
- TikTok** | *Software Engineer Intern* | *C++, Lua, Game Engine, AR/VR* 2022.5 - 2022.8
- **Amazing Engine** | *TikTok's Next-Generation 3D Game Engine for AR/VR Effects*
    - Collaborated in developing TikTok's **3D Game Engine**, which empowers users to create/use interactive **AR/VR** stickers
    - Implement a query-based animation system **Motion Matching** in C++ for realistic and responsive avatar control
    - Developed an **SDK** for **Skeleton Retargeting** in C++ and **Lua**, supporting animation adaptation across character models
    - Integrated the cross-functional team's **Text-to-Animation** algorithm into our game engine using the developed SDK
- University of California, Los Angeles** | *Visiting Student Researcher* | *ML, RL, GNN, FPGA, EDA* 2022.6 - 2022.11
- **GNNDSE** | *An automated design space exploration for automatic FPGA accelerator design* | *Advisor: Prof. Jason Cong*
    - Combined **GNN** with an **ML/RL-based Design Space Exploration** to achieve **FPGA Accelerator Design Automation**
    - Developed a learning-based **Cost Model** with **GNN** as a surrogate of the HLS tool for quick and accurate assessment
    - Optimized DSE by deploying heuristic algorithms such as **Genetic Algorithm** and **Simulated Annealing**
    - Used **Reinforcement-Learning** and **Bandits** for automatic algorithm selection, boosting exploration speed by 11%

## Selected Projects

- Optimized GPU code generation framework for SParse reguLAR Attention** 2023.8 – 2024.7
- Developed SPLAT, an optimized framework for efficient **sparse-MHSA**, targeting moderate sparsity levels
  - Introduced **Affine Compressed Sparse-Row (ACSR)** format for **regular sparsity** patterns in MHSA
  - Engineered advanced **GPU code-generation algorithms** for ACSR, enhancing sparse-MHSA kernel performance
  - Achieved 2.05x and 4.05x speedups over **Triton** and **TVM** kernels with SPLAT implementation
- Remote Car Control System with Real-time 3D Reconstruction** 2023.1 – 2023.5
- Developed a Raspberry Pi robot car with remote control via joysticks, utilizing PID control and STM32 microcontroller
  - Implemented WiFi-based communication for transmission of commands and RGBD images between the car and server
  - Implemented the **SLAM** framework RTAB-Map on the server for real-time **3D reconstruction**, achieving a 10Hz framerate
- Implement A [Game](#) Efficiently on the FPGA Board** 2022.1 - 2022.5
- Ported the game "Doodle Dump" to FPGA with **SystemVerilog**, achieving low power consumption and high efficiency
  - Implemented a SOC with NIOS II in C to manage complex tasks like USB protocol and memory I/O
  - Consumed only 400KB memory, 0.5w power to achieve a 50hz frame rate, won the **Best Design Prize**

## Selected Research

- [SPLAT](#): A framework for optimised GPU code-generation for SParse reguLAR ATtention