

Yipin Guo

 Personal Website
  Google Scholar
  LinkedIn
  guoyipinkevin@gmail.com
  yguo23@nd.edu

Research Interests

Efficient Machine Learning; System for ML; Machine learning algorithm-hardware co-design.

Education

- | | | |
|------------|---|--|
| PhD | University of Notre Dame , Computer Science and Engineering
<ul style="list-style-type: none"> • Advisor: Professor Siddharth Joshi. • Awards: The Arthur J. Schmitt Leadership Fellowship. • GPA: 3.93/4.0. | 09/2025 – 05/2030
South Bend, IN, USA
(Anticipated Graduation) |
| MS | Zhejiang University , Control Science and Engineering
<ul style="list-style-type: none"> • Advisor: Professor Qinyuan Ren • Direction: Pattern Recognition and Intelligent Systems • GPA: 91/100 • TOEFL: 104 (Listening 26; Reading 29; Writing 27; Speaking 22) | 09/2022 – 06/2025
Hangzhou, China |
| BS | Zhejiang University , Automation
<ul style="list-style-type: none"> • GPA: 3.88/4.0 ; Last two years GPA: 3.93/4 • Awards: The Second Prize Scholarship, 2019-2021 Excellent Academic Award | 09/2018 – 06/2022
Hangzhou, China |

Research Experience

University of Notre Dame, Computer Science and Engineering
09/2025 – Now , Supervised by Prof. [Siddharth Joshi](#).

- **SplitZip: Ultra Fast Lossless KV Compression for Disaggregated LLM Serving**
First Author. In Submission.
 - ◊ Ultra-high codec throughput and lossless LLM KV cache compression technology.
 - ◊ Compress high-frequency exponent values of bf16 data using 4-bit fixed-length encoding.
- QTEA: Ternary LLMs with Sparse Residual Salient Weight and By-Column Optimization. **EMNLP 2026.**

Zhejiang University, Control Science and Engineering
01/2022 – 03/2025, Supervised by Prof. [Qinyuan Ren](#)

- **ShiftAddAug: Augment Multiplication-Free Tiny Neural Network with Hybrid Computation**
First Author. Accepted by 2024 CVPR Workshop: *Efficient Deep Learning for Computer Vision*.
 - ◊ Use multiplication to augment the ShiftAdd tiny neural networks during training.
 - ◊ Make ShiftAdd neural networks perform better than mult. counterparts without inference overhead.
- **Efficient Fusion and Task Guided Embedding for End-to-end Autonomous Driving**
First Author. **Best Student Paper Award** of the 13th IEEE Data-Driven Control and Learning Systems Conference.
 - ◊ A lightweight E2E autonomous driving neural network based on Transformer.
 - ◊ 37.6% of the parameters and 8.7% of the calculation with only 0.4% lower driving score.
- **GPTQT: Quantize Large Language Models Twice to Push the Efficiency**
First Author. Accepted by the 11th IEEE International Conference on Cybernetics and Intelligent Systems.
 - ◊ Insight: GPTQ exhibits overfitting during quantization.

- ◇ Two-step quantization: first to relatively high bits, then converted to low-bit binary-coding.
- [Control of Pneumatic Artificial Muscles with SNN-based Cerebellar-like Model](#)
Co-first Author. Accepted by the *13th International Conference on Social Robotics*.
 - ◇ Based on the bionic approach, SNN is used to simulate the cerebellum as a feedforward controller.

Georgia Institute of Technology, School of Computer Science, The EIC Lab (Remote Intern)
01/2023 – 10/2024, Supervised by Prof. [Yingyan \(Celine\) Lin](#), PhD. [Haoran You](#).

- [ShiftAddViT: Mixture of Multiplication Primitives Towards Efficient Vision Transformer](#)
Co-first Author. Accepted by *NeurIPS 2023*.
 - ◇ Linear Additive-Only Attention.
 - ◇ MoE with different calculation. Multiplications for vital tokens and Shift for less crucial ones.
- [ShiftAddLLM: Accelerating Pretrained LLMs via Post-Training Multiplication-Less Reparameterization](#)
Co-first Author. Accepted by *NeurIPS 2024*.
 - ◇ Accelerating pretrained LLMs via a post-training bitwise shift-and-add reparameterization.
 - ◇ Quantization approach to optimize both weight and output activation objectives.
 - ◇ A new mixed and automated bit allocation strategy.
- [5th in 2023 ACM/IEEE TinyML Design Contest at ICCAD](#)
The deployment of the quantized neural network on the MCU.

Projects

- | | |
|--|-----------------|
| Design and development of intelligent handling AGV | 12/2021-07/2023 |
| <ul style="list-style-type: none"> ● AGV controller circuit design. ● MCU software development in STM32 and software architecture design in Linux. | |
| Automatic brake assistance system for intelligent vehicles | 12/2021-06/2022 |
| Zhejiang University, Final Year Project of Bachelor. | |
| <ul style="list-style-type: none"> ● Developing safer AEB strategies and simulating in Matlab. | |

Work Experience

- | | |
|--|-----------------|
| Research Intern , Inco AI. | 06/2026-now |
| Supervised by Prof. Zhijian Liu and Lucas Liebenwein | |
| <ul style="list-style-type: none"> ● Accelerate the generation speed of Gemma-4B to 250 tokens/s on DGX Spark. ● Participate in the training and deployment of VLMs. | |
| Test and Development Engineer , Alibaba DAMO Academy (Intern) | 05/2021-08/2021 |
| Hangzhou, China | |
| <ul style="list-style-type: none"> ● Participate in the testing work of an IOT devices. ● Developing automation script and test AI function. | |

Technologies

Languages: Python, Pytorch, C, C++, Matlab, assembly language, JavaScript.

Design: Lceda, Altium Designer, Solidworks, Multisim.

Equipment: 3D printing, laser cutting machine, welding, oscilloscope, multimeter.