Pengcheng Wang

West Lafayette, US wang4495@purdue.edu +1 765-775-0474

Website LinkedIn Google Scholar GitHub

Education

Ph.D. Candidate - Purdue University **Aug 2019 - Present**

Major: Electrical and Computer Engineering *West Lafayette, IN, US*

M.S. with Excellent Graduate Honor - Tongji University **Sep 2014 - Apr 2017**

Major: Electronic Science and Technology, Minor: Green Economy and Sustainable Development *Shanghai, China*

B.E. with Excellent Graduate Honor - Tongji University **Sep 2010 - Jun 2014**

Major: Electronic Science and Technology *Shanghai, China*

Research and Work Experience

Machine Learning Engineer Intern, EmbodyX **Sep 2025 - Present**

Tools: Python, PyTorch, PEFT, Prismatic VLMs, VLM Evaluation, DeepSpeed, HuggingFace, Docker, GIT, SLURM

- Applying Machine Learning Systems optimization to accelerate LLM and VLM training and inference for foundation model development
- Exploring model compression and related techniques to enhance efficiency and scalability for robotic systems
- Enabling practical deployment of large models in real-world environments by ensuring robust system performance

Adaptive Vision-Language Model (VLM) under Resource Constraints **Jan 2025 - Present**

Tools: Python, PyTorch, LLaVA, DeepSpeed, lmms-eval, HuggingFace, Docker, GIT, SLURM

- Characterizing and exploiting the inference design space of adaptive vision-language model (VLM) systems across token, width, and depth dimensions.
- Developing a noise-resilient and compute-adaptive VLM that meets specific computational budgets while maximizing inference accuracy under noisy inputs.
- Designing a learning-based scheduler that is both noise- and latency-aware, dynamically selecting execution paths for the base VLM during inference.

Software Engineer Intern - AI Toolchain, Sunlune **Jan 2025 - Aug 2025**

Project: Software Toolchain for AI Chip. Tools: Python, C/C++, PyTorch, Pytest, Gem5, Docker, GIT, OpenCL

- Developed and validated Kernel, Runtime, and Driver frameworks for AI accelerators by testing kernel functions, integrating a Hydra configuration system, and evaluating performance on simulation platforms.
- Integrated LLM kernels and optimized runtime workflows to enable efficient inference and validation of Llama-family models on AI accelerators.
- Performed feature testing, performance tuning, and cross-platform debugging to resolve bottlenecks and improve runtime and kernel execution efficiency.

Content-Aware Adaptive 3D Object Detection Systems for Embedded GPUs **Aug 2022 - Jan 2025**

Tools: Python, PyTorch, TensorFlow, OpenPCDet, MMDetection, Docker, GIT, Anaconda, SLURM. Datasets: Waymo, nuScenes, KITTI

- Developed the first adaptive 3D object detection system for LiDAR point cloud data, capable of meeting runtime latency targets ranging from 35 to 200 ms while maximizing accuracy on NVIDIA embedded GPUs.
- Conducted a systematic comparison of 2D and 3D object detection models, highlighting key components unique to 3D pipelines—such as voxelization, voxel feature encoding, and 3D spatial feature extraction.
- Introduced five adjustable control knobs to enable multiple execution branches, leveraging the interplay between system components including the 3D encoder, CNN backbone, and detection head.
- Deployed on NVIDIA Jetson AGX Orin and Xavier; demonstrated 2–5% higher mean Average Precision than baselines, achieving over 20 FPS on Xavier and significantly outperforming models such as CenterPoint and DSVT.

Generative AI Model Intern, Sunlune **May 2024 - Jan 2025**

Project: GAI for Chip Design. Tools: Python, C, Perl, PyTorch, TensorFlow, DRiLLS, yosys, ABC, OpenSTA, dc_shell, Docker, GIT.

Datasets: OpenABC, EPFL Combinational Benchmark Suite, ISCAS High-Level Models

- Developed AI-driven design flows for high-performance digital ASICs, achieving significant reductions in design time and improvements in power, performance, and area (PPA).
- Designed deep reinforcement learning models for logic synthesis and technology mapping, yielding a 10% improvement in delay-power product and enhanced scalability across diverse designs.
- Collaborated with IC design engineers to encode domain expertise into AI models by integrating reinforcement learning with human feedback (RLHF) for more adaptive and explainable optimization.
- Implemented heuristic-based post-processing techniques, leading to substantial reductions in both power consumption and critical path delay.

Adaptive and Efficient Video Object Detection Systems on Edge Devices **Aug 2019 - Dec 2022**

Tools: Python, CUDA C/C++, PyTorch, TensorFlow, Detectron2, MMDetection, MMTracking, GIT, Anaconda. Datasets: ILSVRC

- Performed comprehensive energy benchmarking on 10+ object detection models (e.g., YOLO, EfficientDet, Faster R-CNN) on an embedded testbed under varying power constraints.
- Evaluated latency and accuracy of 20+ classification and detection models on the ILSVRC Video dataset, identifying optimal configurations under limited compute resources.
- Developed a synthetic 3D contention generator (CPU, memory bandwidth, GPU) to simulate controllable workloads in resource-constrained environments.
- Established and maintained a heterogeneous embedded testbed with NVIDIA Jetson platforms (Nano, TX2, Xavier, Orin), and built a robust software stack (JetPack, Docker, OpenCV, PyTorch, TensorFlow, Python) to support algorithm development and evaluation.

Selected Publications

- **Pengcheng Wang**, ZhuoMing Liu, Shayok Bagchi, Ran Xu, Saurabh Bagchi, Yin Li, and Somali Chaterji. "Agile3D: Adaptive Contention- and Content-Aware 3D Object Detection for Embedded GPUs" *The 23rd ACM International Conference on Mobile Systems, Applications, and Services (MobiSys) 2025*.
- **Pengcheng Wang**, Shayok Bagchi, Yin Li, and Somali Chaterji. "Adapt3D: Adaptive 3D Object Detection System for Embedded GPUs." *The 17th International Conference on COMmunication Systems & NETworkS (COMSNETS) 2025*.
- Lee, Jayoung, **Pengcheng Wang**, Ran Xu, Sarthak Jain, Venkat Dasari, Noah Weston, Yin Li, Saurabh Bagchi, and Somali Chaterji. "Virtuoso: Energy-and Latency-Aware Streamlining of Streaming Videos on SOCs." *ACM Transactions on Design Automation of Electronic Systems (TODAES) 2023*.
- Xu, Ran, Jayoung Lee, **Pengcheng Wang**, Saurabh Bagchi, Yin Li, and Somali Chaterji. "LiteReconfig: cost and content aware reconfiguration of video object detection systems for mobile GPUs." *In Proceedings of the Seventeenth European Conference on Computer Systems (EuroSys) 2022*.
- Xu, Ran, Rakesh Kumar, **Pengcheng Wang**, Peter Bai, Ganga Meghanath, Somali Chaterji, Subrata Mitra, and Saurabh Bagchi. "ApproxNet: Content and contention-aware video object classification system for embedded clients." *ACM Transactions on Sensor Networks (TOSN) 2021*.
- **Pengcheng Wang**, Jayoung Lee, Ran Xu, Venkat Dasari, Noah Weston, Yin Li, Saurabh Bagchi, and Somali Chaterji. "Benchmarking video object detection systems on embedded devices under resource contention." *In Proceedings of the 5th International Workshop on Embedded and Mobile Deep Learning (EMDL) 2021*.
- Xu, Ran, Chen-lin Zhang, **Pengcheng Wang**, Jayoung Lee, Subrata Mitra, Somali Chaterji, Yin Li, and Saurabh Bagchi. "ApproxDet: content and contention-aware approximate object detection for mobiles." *In Proceedings of the 18th Conference on Embedded Networked Sensor Systems (SenSys) 2020*.
- Shankar, Karthick, **Pengcheng Wang**, Ran Xu, Ashraf Mahgoub, and Somali Chaterji. "Janus: Benchmarking commercial and open-source cloud and edge platforms for object and anomaly detection workloads." *In 2020 IEEE 13th International Conference on Cloud Computing (CLOUD) 2020*.

Certifications

- Building RAG Agents with LLMs
- Introduction to Transformer-Based Natural Language Processing
- Generative AI with Diffusion Models
- Certificate of Excellence for Hugging Face Deep Reinforcement Learning Course
- Building Real-Time Video AI Applications
- Building Video AI Applications at the Edge on Jetson Nano
- Disaster Risk Monitoring Using Satellite Imagery
- Fundamentals of Accelerated Computing with CUDA Python
- Fundamentals of Accelerated Computing with CUDA C/C++

Honors and Awards

- Selected Mentee, GradBridge Program (Purdue & UC Berkeley), 2026
- Outstanding International Student Alumni Award, Tongji Foundation, 2025
- Dependable Computing Systems Laboratory Group Champion, Purdue University, 2025
- NSF Workshop on Grand Challenges in Resilience Poster Winner, Purdue University, 2024
- NSF Workshop on Grand Challenges in Resilience Poster Winner, Purdue University, 2023
- Excellent Graduate, Tongji University, 2017
- National Scholarship for Graduate Students, Ministry of Education of China, 2016
- Excellent Master Scholarship, Tongji University, 2015
- Excellent Graduate, Tongji University, 2014
- Scholarship for Good Academic Performance, Tongji University, 2013