

Pengcheng Wang

wang4495@purdue.edu

[Website](#) [LinkedIn](#) [Google Scholar](#) [GitHub](#)

SUMMARY

AI systems researcher developing efficient and adaptive foundation-model systems for real-world applications. My research combines algorithm-system co-design, runtime scheduling, and model optimization for heterogeneous hardware to improve LLM/VLM inference and multimodal perception across edge GPUs, server GPUs, and AI accelerators. I focus on preserving task performance under constraints on latency, throughput, serving cost, and energy. My work has led to publications in top AI and systems venues, complemented by industry experience in foundation-model optimization and AI accelerator software.

EDUCATION

Ph.D. Candidate - Purdue University

Electrical and Computer Engineering. Advisors: [Prof. Somali Chaterji](#), [Prof. Saurabh Bagchi](#)

Aug 2019 - Present

West Lafayette, IN, US

M.S., Tongji University

Electronic Science and Technology. Excellent Graduate Honor. Advisor: [Prof. Meisong Tong](#)

Sep 2014 - Apr 2017

Shanghai, China

B.E., Tongji University

Electronic Science and Technology. Excellent Graduate Honor. Advisor: [Prof. Meisong Tong](#)

Sep 2010 - Jun 2014

Shanghai, China

EXPERIENCE

Machine Learning Engineer Intern, [EmbodimentX](#)

Foundation model optimization for multimodal AI systems

- Developing systems-level methods to reduce the training and inference cost of LLMs and VLMs for multimodal foundation-model workloads.
- Evaluating model compression and parameter-efficient fine-tuning strategies to preserve task performance under compute, memory, and latency budgets.
- Translating research prototypes into deployable model-system components, with emphasis on robust inference behavior in real-world deployment settings.

Sep 2025 - Present

West Lafayette, IN

Adaptive Vision-Language Models (VLMs) under Resource Constraints

Purdue University. Advisors: [Prof. Somali Chaterji](#), [Prof. Yin Li](#), [Prof. Saurabh Bagchi](#)

- Introduced a joint token-compute adaptation framework for multimodal LLMs, reducing inference cost while preserving task accuracy.
- Developing a noise-resilient, compute-adaptive VLM framework that satisfies specified compute budgets while maximizing accuracy under degraded inputs.
- Formulating a learning-based inference scheduler that selects execution paths based on input content, input quality, and latency constraints.

Jan 2025 - Present

Software Engineer Intern - AI Accelerator Toolchain, [Sunlune Corp.](#)

AI chip software infrastructure

- Developed and validated kernel, runtime, and driver frameworks for domain-specific AI accelerators using cycle-accurate simulation platforms.
- Integrated LLM inference kernels and optimized runtime scheduling to support efficient execution of Llama-family models on custom AI hardware.
- Performed systematic performance profiling and cross-platform debugging to identify and remove kernel-level bottlenecks on novel architectures.

Jan 2025 - Aug 2025

Santa Clara, CA

Generative AI for Chip Design, [Sunlune Corp.](#)

AI-driven electronic design automation

- Developed AI-driven design automation flows for high-performance digital ASICs, achieving significant reductions in design iteration time and improvements in PPA metrics.
- Designed deep reinforcement learning models for logic synthesis and technology mapping, yielding a 10% improvement in delay-power product with enhanced scalability across diverse circuits.
- Integrated RLHF to encode IC design domain expertise into optimization models, enabling more adaptive and interpretable design space exploration.
- Developed heuristic-based post-processing techniques that produced substantial reductions in both power consumption and critical path delay.

May 2024 - Jan 2025

Santa Clara, CA

Content-Aware Adaptive 3D Object Detection for Embedded GPUs

Purdue University. Advisors: [Prof. Somali Chaterji](#), [Prof. Yin Li](#), [Prof. Saurabh Bagchi](#)

- Proposed the first adaptive 3D object detection system for LiDAR point cloud data, capable of meeting runtime latency targets (35–200 ms) while maximizing accuracy on NVIDIA embedded GPUs.
- Conducted a systematic architectural analysis of 3D detection pipelines, identifying key computational bottlenecks in voxelization, voxel feature encoding, and 3D spatial feature extraction.

Aug 2022 - Jan 2025

- Introduced five adjustable control knobs spanning the 3D encoder, CNN backbone, and detection head, enabling fine-grained accuracy-latency trade-offs through multiple execution branches.
- Deployed and evaluated on NVIDIA Jetson AGX Orin and Xavier; demonstrated 2–5% higher mAP than baselines at over 20 FPS, significantly outperforming CenterPoint and DSVT.

Adaptive and Efficient Video Object Detection on Edge Devices

Aug 2019 - Dec 2022

Purdue University. Advisors: [Prof. Somali Chaterji](#), [Prof. Yin Li](#), [Prof. Saurabh Bagchi](#)

- Benchmarked 10+ object detection models (YOLO, EfficientDet, Faster R-CNN) on embedded platforms under varying power and thermal constraints.
- Developed a synthetic 3D resource contention generator (CPU, memory, GPU) for controlled evaluation of model robustness in multi-tenant edge settings.
- Maintained a heterogeneous embedded testbed (NVIDIA Jetson Orin, Xavier, TX2, Nano) with full software stack for reproducible evaluation.

SELECTED PUBLICATIONS

- **Pengcheng Wang**, Zhiqian Wang, Jayoung Lee, Zhuoyan Xu, Ran Xu, Saurabh Bagchi, Yin Li, and Somali Chaterji. “[Look Less, Think Faster: Joint Token-Compute Adaptation for Multimodal LLMs.](#)” *European Conference on Computer Vision (ECCV)*, 2026.
- **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.
- 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.

SKILLS

Programming Languages: Python, C/C++, CUDA, Bash, $\LaTeX$, Markdown.

AI/ML Frameworks: PyTorch, TensorFlow, HuggingFace Transformers (LLMs and vision/multimodal models).

LLM/VLM Training & Fine-tuning: DeepSpeed, PyTorch FSDP, HuggingFace PEFT, LLaVA / Qwen-VL.

Inference & Deployment: ONNX, TensorRT, NVIDIA Jetson (Thor/Orin/Xavier/TX2/Nano).

Profiling & Experiment Tracking: PyTorch Profiler, NVIDIA Nsight, Weights & Biases (W&B).

Tools & Platforms: Git, SLURM, Docker, AWS, Linux, CMake.

SELECTED HONORS AND AWARDS

- NVIDIA Academic Grant Program, 2026.
- 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.

SELECTED PROFESSIONAL DEVELOPMENT

NVIDIA Deep Learning Institute: [Building RAG Agents with LLMs](#), [Introduction to Transformer-Based Natural Language Processing](#), [Generative AI with Diffusion Models](#), [Fundamentals of Accelerated Computing with CUDA Python](#), [Fundamentals of Accelerated Computing with CUDA C/C++](#), [Building Real-Time Video AI Applications](#).

Hugging Face: [Certificate of Excellence for Deep Reinforcement Learning](#).