Pengcheng Wang

West Lafayette, US wang4495@purdue.edu
Website LinkedIn Google Scholar GitHub

Summary

ML systems engineer focused on efficient and adaptive inference of large language models (LLMs) and vision-language models (VLMs) across server GPUs, AI accelerators, and embedded edge platforms. Builds compute-adaptive VLMs through token compression, model adaptation, and runtime scheduling, with industry experience deploying foundation models under real-world latency, energy, and resource constraints. Publications at top systems venues including MobiSys, EuroSys, SenSys, and TODAES.

Education

Ph.D. Candidate - Purdue University **Aug 2019 - Present**
Electrical and Computer Engineering. Advisors: Prof. Somali Chaterji, Prof. Saurabh Bagchi *West Lafayette, IN, US*

M.S., Tongji University **Sep 2014 - Apr 2017**
Electronic Science and Technology. Excellent Graduate Honor. Advisor: Prof. Meisong Tong *Shanghai, China*

B.E., Tongji University **Sep 2010 - Jun 2014**
Electronic Science and Technology. Excellent Graduate Honor. Advisor: Prof. Meisong Tong *Shanghai, China*

Experience

Machine Learning Engineer Intern, EmbodyX **Sep 2025 - Present**
Foundation model optimization for robotic systems *West Lafayette, IN*

- Investigating systems-level optimization techniques to accelerate training and inference of LLMs and VLMs for foundation model development.
- Exploring model compression and parameter-efficient fine-tuning methods for scalable deployment of foundation models on compute-constrained platforms.
- Designing system-level solutions that ensure robust inference performance of large-scale models in real-world deployment environments.

Adaptive Vision-Language Models (VLMs) under Resource Constraints **Jan 2025 - Present**
Purdue University. Advisors: Prof. Somali Chaterji, Prof. Yin Li, Prof. Saurabh Bagchi

- Characterizing the inference design space of VLMs across token, width, and depth dimensions to expose efficiency–accuracy trade-offs.
- Building a noise-resilient and compute-adaptive VLM framework that meets specified compute budgets while maximizing accuracy under degraded inputs.
- Designing a learning-based, noise- and latency-aware inference scheduler that dynamically selects execution paths per sample.

Software Engineer Intern – AI Accelerator Toolchain, Sunlune Corp. **Jan 2025 - Aug 2025**
AI chip software infrastructure *Santa Clara, CA*

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

Generative AI for Chip Design, Sunlune Corp. **May 2024 - Jan 2025**
AI-driven electronic design automation *Santa Clara, CA*

- 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.

Content-Aware Adaptive 3D Object Detection for Embedded GPUs **Aug 2022 - Jan 2025**
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.
- 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 Nano, TX2, Xavier, Orin) with full software stack for reproducible 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*.

Skills

Programming Languages: Python, C/C++, CUDA, Bash, LaTeX, Markdown.
AI/ML Frameworks: PyTorch, TensorFlow, HuggingFace Transformers (text and vision/multimodal).
LLM/VLM Training & Fine-tuning: DeepSpeed, PyTorch FSDP, HuggingFace PEFT, LLaVA / Qwen-VL.
Inference & Deployment: ONNX, TensorRT, NVIDIA Jetson (Nano/TX2/Xavier/Orin).
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.