Yuchen Wang

(217)480-2166 | yuchenwang0303@gmail.com
Yuchen | yuchenwang3

Education

- **University of Illinois Urbana-Champaign (UIUC)** *2025 – 2027 (expected)*
  M.S. in Computer Science *GPA: 4.0*
- **Peking University** *2021 – 2025*
  B.S. in Intelligence Science and Technology *GPA: 3.73 (Top 20%)*

Skills

- **Programming:** Python, C, C++, Bash, SQL
- **Machine Learning:** PyTorch, Hugging Face Transformers, scikit-learn, NumPy, Pandas
- **Generative AI & LLMs:** LLM fine-tuning, RLHF / GRPO, Diffusion Models, ControlNet, RAG
- **MLOps:** ML pipelines, model serving, performance monitoring, profiling, inference optimization
- **Systems & Deployment:** vLLM, Megatron-LM, SGLang, CUDA, Docker, Linux, CI/CD concepts
- **Tools:** Git, GitHub, Logging & Evaluation, Nsight

Internship Experience

- **Freedo Technology** *Summer 2025*
  Research Intern Beijing, China
  - Built an end-to-end **generative AI pipeline** using **depth-conditioned ControlNet** for 3D building reconstruction from noisy point clouds, covering data preprocessing, model inference, and post-processing.
  - Designed depth and normal map conditioning to improve robustness under heavy noise, preserving structural edges and planar surfaces.
  - Improved reconstruction **geometric fidelity from ~30% to ~80%** on representative datasets.

Project Experience

- **Dynamic Prefill Optimization via Adaptive Online Packing** *Sep 2025 – Dec 2025*
  Workload-aware batching and request packing for low-latency inference Tools: Python, vLLM, Profiling & Evaluation
  - Designed an **AIMD-based dynamic batching controller** that adapts prefill trigger thresholds using real-time **p95 TTFT** feedback, with burst overrides for non-stationary traffic.
  - Implemented **length-aware packing** for heterogeneous prompts (greedy baseline + **DP-based optimal packing**) to reduce padding waste and improve token utilization.
  - Evaluated on **production traces (DynamoLLM)**; achieved **up to 20% lower end-to-end TTFT** compared to Prepacking under conversational workloads.
- **FlashAttention v1 and CUDA Kernel Optimization for GPT-2 Inference** *Sep 2025 – Dec 2025*
  Memory-efficient attention and CUDA kernel optimization Tools: CUDA, cuBLAS, Nsight
  - Built an optimized **GPT-2 inference pipeline** with custom CUDA kernels for key operators, focusing on end-to-end throughput and latency.
  - Implemented **FlashAttention v1** (tiled + online softmax + blockwise loop) to avoid materializing the full attention matrix; reduced **HBM traffic from 1.57MB to 0.15MB** and improved arithmetic intensity by **10.3×**, while maintaining numerical stability (max error < 1e-5).
  - Reduced kernel launch overhead by fusing attention-related kernels (3–4 launches per layer → fewer launches), saving **50% kernel launch cost** in the attention path.
  - Conducted profiling-driven tuning with **Nsight** and configuration sweeps; achieved **up to 9% speedup** over baseline in the best configuration.
  - Implemented kernel-level operator fusion (e.g., GEMM + Bias + GELU, fused QKV projection) to reduce intermediate tensors and global memory round-trips.
- **Reinforcement Learning for Legal Reasoning on Multi-Choice QA** *Feb 2025 – May 2025*
  Advisor: Prof. Yansong Feng, Peking University Tools: PyTorch, Qwen
  - Developed a hybrid training framework (Zero-RL → SFT with distilled CoT → GRPO) to enhance LLM reasoning on Chinese bar-exam multi-choice case-analysis questions.
  - Achieved **57.6% accuracy** on benchmark test set, surpassing SFT-only and RL-only baselines.
  - Enabled structured legal reasoning with precise statute citation and detailed option-by-option analysis.
- **Interactive World Model and Variable-Length Video Generation** *Oct 2025 – Present*
  Advisor: Prof. Fan Lai, University of Illinois Urbana–Champaign Tools: PyTorch, Diffusion Transformers, LLMs
  - Engineering an end-to-end **xDiT experimentation pipeline** for training and inference (config-driven runs, logging, checkpointing, and evaluation hooks) to support rapid iteration.
  - Implementing **variable-length generation** support (data pipeline + length conditioning/masking + sampling utilities) and setting up baselines/ablations for long-horizon temporal consistency.
  - Prototyping **retrieval-and-reuse** and **speculative chunking** modules to reduce interactive latency; benchmarking latency/VRAM/quality trade-offs on representative prompts.

Honors and Awards

- **Zhi Class Scholarship** *Peking University* *2023, 2024*
- **Merit Student** *Peking University* *2023*
- **First Prize, Provincial Chinese Mathematical Olympiad (CMO)** *Chinese Mathematical Olympiad Committee* *2020*