Yuchen Wang

(217)480-2166 | yuchenwang0303@gmail.com
LinkedIn | GitHub | Portfolio

Education

- **University of Illinois Urbana-Champaign (UIUC)** *2025 – 2027 (expected)*
  M.S. in Computer Science Champaign, IL, USA
- **Peking University** *2021 – 2025*
  B.S. in Intelligence Science and Technology *Zhi Class (Elite Program), Top 10%* Beijing, China

Internship Experience

- **Alibaba Group (U.S.)** *May 2026 – Present*
  Research Intern — Agentic LLM Post-Training Sunnyvale, CA, USA
  - Developed Occamy-1.0, a 35B-A3B co-work model continued from Qwen3.6-35B-A3B through end-to-end post-training with full-parameter SFT, uniform model soup, and GRPO/SAO reinforcement learning.
  - Built a verifier-gated, multi-harness data and training pipeline with immutable provenance, token-exact proxy/TiTO replay, state reconstruction, episode credit, and quarantine gates; achieved **38.6% fewer tokens/trajectory and 36.1% lower wall time**, with timeouts falling from 10.89% to 4.69% and invalid calls from 3.28% to 1.68%.
  - Raised Combined ClawEval T/C Strict Pass@1/3 from 65.16/73.87% to **77.39/85.93%**, with 46.9 on WildClawBench and 27.6% strict / 69.1% partial on AutomationBench 1.0.6.

- **Freedo Technology** *Summer 2025*
  Research Intern — 3D Vision and Generative AI Beijing, China
  - Built an end-to-end depth-conditioned ControlNet pipeline for 3D building reconstruction from noisy point clouds; designed depth/normal conditioning to preserve structural edges and planar surfaces, improving geometric fidelity from **32.7% to 85.3%**.

Open-Source Systems Engineering

- **Training reliability:** Muon numerical-stability fixes in NVIDIA NeMo, Mamba+attention+MoE recompute/runtime fixes in vLLM/Vime, and four ms-swift patches spanning sequence packing, NCCL startup OOM prevention, and optimizer correctness.
- **Long-context & RL systems:** Megatron-LM #5396 fuses GatedDeltaNet Q/K normalization for 128K SFT; #5463 adds selective Mamba recompute validated on 8×B200 (24K BF16 completed where the baseline OOMed); NeMo RL #2962 sanitizes non-finite rollout logprobs; SGLang #31621 hardens hybrid-model weight reloads.

Papers & Projects

- **CineFlow: Dependency-Driven Parallel Video Generation** *[Paper]* *Oct 2025 – May 2026*
  Advisor: Prof. Fan Lai, University of Illinois Urbana–Champaign Tools: PyTorch, xDiT, Diffusion Transformers
  - Evaluated semantic-dependency scheduling and trajectory-aware fusion on Wan2.2-5B, CogVideoX-5B, and HunyuanVideo across 8×H100s: **1.7–5.5× end-to-end speedup and 5.4–17.3% higher VBench overall**, and 1.30–2.02× lower P90 latency; ablations measured 6.7–14.6% higher visual coherence and nearly 7% more throughput from online dependency pruning.

- **Dynamic Prefill Optimization for LLM Inference** *[Report]* *Aug 2025 – Jan 2026*
  Adaptive Online Packing — University of Illinois Urbana–Champaign Tools: Python, CUDA, vLLM, Profiling
  - Built an AIMD batching controller with real-time p95 TTFT feedback and greedy/DP length-aware packing before continuous batching; achieved up to **20% lower TTFT** on production-style DynamoLLM traces.

- **FlashAttention-style CUDA Kernels for GPT-2** *[Report]* *Aug 2025 – Dec 2025*
  University of Illinois Urbana–Champaign Tools: CUDA, C++, Nsight Compute/Systems
  - Implemented tiled attention with online softmax and kernel fusion, raising arithmetic intensity from 15 to 157 FLOP/byte and delivering **∼10× lower HBM traffic and up to ∼9% end-to-end speedup**.

- **RL for Legal Reasoning on Multi-Choice QA** *[Thesis]* *Jan 2025 – Mar 2025*
  Peking University Tools: PyTorch, Qwen, GRPO
  - Built a Zero-RL → distilled-CoT SFT → GRPO pipeline for structured legal reasoning, reaching **57.6% accuracy** and outperforming SFT-only and RL-only baselines.

Skills

Python, C++, CUDA, PyTorch, 3D vision, point clouds, ControlNet, diffusion transformers, Megatron-LM, NeMo-RL, vLLM, SGLang, distributed training, profiling