Wenxuan Xu

(213) 823-0401 | lilmeep727@gmail.com| Wen-xuan-Xu | Google Scholar wen-xuan-xu.github.io Wenxuan Xu

Education

Dartmouth College 2024/09 – 2026/06 (Expected)
MS, Computer Science with Concentration on Digital Arts Hanover, NH, USA
University of Liverpool (Xi'an Jiaotong-Liverpool University) 2020/09 – 2024/07
BS, Computer Science (GPA 3.78/4.00, First Class Honors, Dual Degree) Liverpool, UK | Suzhou, China

Technical Skills

Programming & HPC:
- Proficient in **C/C++** and **Python**. Expert in **CUDA** programming with deep knowledge of GPU memory hierarchy (Global/Shared/Register) and hardware architectures (**Hopper/Ampere**).
- Skilled in kernel optimization techniques including **Tiling**, **Loop Unrolling**, **Thread Coarsening**, and resolving **Bank Conflicts**.

Inference Systems & Architecture:
- Source-level understanding of **SGLang** core mechanisms: **Radix Attention**, **Continuous Batching**, **Chunked Prefill**, and **Speculative Decoding**.
- Hands-on experience profiling and tuning kernels on **NVIDIA H200/B200** clusters (contributed benchmarks to the SGlang community).

Distributed Systems & Tools:
- Proficient in distributed strategies (**TP/PP/DP/EP**) and **NCCL** communication optimization.
- Advanced profiling skills using **Nsight Compute (ncu)** and **Nsight Systems (nsys)** for kernel and system-level analysis.
- Experienced with **Linux** system programming, **CMake**, **Docker/Kubernetes**, and Git workflows.

Research Experience

HealthX Lab, *Dartmouth College* | Research Assistant *2025/03 - 2025/12*
LENS–Multimodal LLM for Clinical Mental-Health Narratives | Advisor: Prof. Andrew Campbell
- Engineered a data pipeline for a large-scale MDD study (**51k samples**). Developed a **multi-agent "LLM-as-a-judge"** framework to validate synthetic narratives, curating a high-quality instruction-tuning dataset of **>150k multimodal QA pairs**.
- Architected LENS, a **time-series multimodal LLM**. Designed a **patch-based MLP projection layer** to map sensor data directly into the **Qwen2.5** embedding space, enabling efficient **end-to-end reasoning** over raw sensor streams and text.
- Surpassed text-serialization baselines in narrative quality and symptom alignment. Achieved **clinician-level performance** comparable to larger VLMs while **reducing token consumption by 10x**.
- Executed a **two-stage curriculum training** strategy (encoder alignment followed by SFT) on **H200 clusters**, ensuring robust multimodal alignment across variable-length temporal sequences.

Pervasive HCI Group, *Tsinghua University* | Research Intern
FIT-AWE Lab, *the Hong Kong University of Science and Technology (Guangzhou)* | Research Intern *2022/07 – 2024/09*
Multimodal LSTM for Ray Pointer Prediction in VR | Advisor: Prof. Hai-Ning Liang and Prof. Yuntao Wang
- Built a VR study platform in Unity + Meta Quest Pro, recording 72 k head-, eye-, and hand-tracking sequences at 90 Hz from bare-hand ray-pointing tasks.
- Trained a **tri-modal stacked LSTM** on velocity- and distance-time-series to predict ray-landing poses, added gaze-driven early-stage prediction and cross-user generalization, and ran **head / hand / eye ablation tests** to quantify each modality's role in human motor control.
- Reached **1.9**× lower angular error and **3.5**× higher hit-rate at 40 % of the movement phase, outperforming kinematic baselines; results published at **IEEE VR '25**.

Projects

LiteInfer – High-Performance LLM Inference Engine (C++/CUDA) *2025/07 – 2025/10*
- Architected a lightweight inference framework supporting Llama 3.2 and Qwen2.5. Implemented **Continuous Batching** and **PagedAttention** to mitigate memory fragmentation, increasing KV-cache throughput by **>30%** under high-concurrency workloads.
- **Kernel Optimization**: Engineered custom CUDA kernels for FlashAttention-2, RMSNorm, and RoPE. Leveraged **Nsight Compute** to resolve bank conflicts and optimize warp occupancy. Achieved a **28% latency reduction** in Attention ops and boosted system throughput from 82 to **112 tokens/s** (8K context) via **Shared Memory tiling** and Tensor Core pipelining.
- **Quantization & System Integration**: Developed an **Int8/AWQ** group-wise quantization scheme with fused de-quantization kernels. Reduced memory footprint by **50%** with <1% accuracy loss and minimized kernel launch overhead using **CUDA Graphs**.

Teacher-Guided Token Re-Weighting for SFT Reasoning (PyTorch) *2025/09 – Present*
- Proposed a dynamic re-weighting SFT method using Teacher-Student logits divergence. Constructed a dual-forward pipeline where **Qwen2.5-32B** guides a **1.5B student**, adaptively amplifying gradients for critical reasoning tokens to prevent "step loss" in long CoT paths.
- **Training Infra Optimization**: Implemented efficient fine-tuning on **8×H200** clusters using **FSDP + LoRA**. Optimized training scripts with Gradient Checkpointing and mixed-precision training, achieving **>90% GPU memory utilization**.
- **Results**: Achieved a **20% Pass@1 improvement** on MATH-500 and AIME benchmarks. Significantly enhanced the stability and logical coherence of Chain-of-Thought reasoning compared to standard SFT baselines.

Publications

1. **[ACL ARR' 26 (Pre-Print)] Wenxuan Xu***, Arvind Pillai*, Subigya Nepal, Amanda C Collins, Daniel M Mackin, Michael V Heinz, Tess Z Griffin, Nicholas C Jacobson, Andrew Campbell. *"LENS: LLM-Enabled Narrative Synthesis for Mental Health by Aligning Multimodal Sensing with Language Models"*
2. **[IEEE VR' 25] Wenxuan Xu**, Yushi Wei, Xuning Hu, Wolfgang Stuerzlinger, Yuntao Wang, Hai-Ning Liang. "*Predicting Ray Pointer Landing Poses in VR Using Multimodal LSTM-Based Neural Networks"*
3. **[IEEE VR' 25]** Xuning Hu*, **Wenxuan Xu***, Yushi Wei, Zhang Hao, Jin Huang, Hai-Ning Liang. *"Optimizing Moving Target Selection in VR by Integrating Proximity-Based Feedback Types and Modalities"* (**Co-first author**)
4. **[ISMAR' 24]** Xuning Hu, Xinan Yan, Yushi Wei, **Wenxuan Xu**, Yue Li, Yue Liu, Hai-Ning Liang. *"Exploring the Effects of Spatial Constraints and Curvature for 3D Piloting in Virtual Environments"*
5. **[IEEE TVCG' 26]** Yifan Qi*, Xuning Hu*, Xinan Yan, **Wexuan Xu**, Hao Zhang, Hai-Ning Liang, Jin Huang. *"Exploring Freehand-Based Selection Techniques of Polyhedron Faces in VR Environments"*