Weili Xu (徐纬立)

weilixu2@illinois.edu
weili-0234.github.io
github.com/weili-0234
linkedin.com/in/weili-xu-2a05662a7

306 N Wright St.
Urbana, IL, 61801
Dept. of Electrical & Computer Engineering
University of Illinois Urbana-Champaign

Education

University of Illinois Urbana-Champaign *Aug. 2023 – present*
B.S. in Computer Engineering, GPA: 4.0/4.0

Ongoing Coursework:
Applied Parallel Programming, Computer Systems Engineering, Deep Generative Models

Zhejiang University *Aug. 2023 – present*
B.S. in Computer Engineering, GPA: 4.20/4.30, Rank: 6/70

Selected Coursework:
Data Structure (A+), Discrete Mathematics (A+), Linear Algebra (A+), Computer Systems & Programming (A)

Research

- **Efficient LLMs.** Transformer LLMs require a KV-cache that scales linearly with input length, creating system bottlenecks for long-context modeling. I aim to tackle this challenge with algorithmic innovations such as sparse attention and linear attention that reduce wall-clock runtime with hardware-aware implementation.
- **Video understanding.** A key challenge in LLM-based video understanding is to maintain memory efficiency during training and inference while managing GPU memory costs that limit input video length. My previous work addresses this by incorporating efficient Linear Attention LLM backbones into LLaVA-like video LLMs.

Research Interest: Efficient LLMs and their Applications in Long Context (e.g. Agents and Long Videos)

Publication

* denotes equal contribution.

Video Understanding

1 **AuroraLong: Bringing RNNs Back to Efficient Open-Ended Video Understanding**
Weili Xu*, Enxin Song*, Wenhao Chai, Xuexiang Wen, Tian Ye, Gaoang Wang
Accepted as Poster by International Conference on Computer Vision (ICCV), 2025. arXiv: 2507.02591
In this work, we propose a hybrid Multimodal LLM (MLLM) based on RWKV that efficiently handles hour-long videos on a single NVIDIA 3090 GPU, achieving comparable performance to its Transformer counterparts.

2 **Video-MMLU: A massive multi-discipline lecture understanding benchmark**
Enxin Song*, Wenhao Chai*, **Weili Xu***, Jianwen Xie, Yuxuan Liu, Gaoang Wang
To appeaar as Oral presentation in KnowledgeMR Workshop, ICCV 2025. arXiv: 2504.14693
Video-MMLU is a massive benchmark designed to evaluate the multimodal reasoning capabilities of MLLMs in understanding Multi-Discipline Lecture Videos.

Technical Blog

Efficient LLM Architecture

1 **View Transformer Layers from Online Optimization Perspective**
Wenhao Chai, **Weili Xu***. Webpage
In this blog, we introduce Mesa Layer and Test-time Training (TTT), and show that standard autoregressive Transformers can implicitly perform gradient-based online optimization during inference.

Working Experience

Undergraduate Teaching Assistant for ECE 120 (Introduction to Computing) Fall 2024
Undergraduate Teaching Assistant for ECE 220 (Computer Systems & Programming) Spring 2025