Yuxuan Wang

(+86)18157865833 (wechat) • yxwang1215@gmail.com

Education

Xidian University **Xi'an, China**
Undergraduate in Software Engineering *Sep. 2023 – June. 2027*

- **GPA: 4.0/4.0** **CET-6:** 591
- **Rank: 1/1342** (Computer Category, 2023-2024) **1/335** (Software Engineering)
- **Core Courses:** Physics (I)(99) Introduction of Computer and Program Design(98), Fundamentals of Circuits and Electronic Techniques(98), Advanced Mathematics A(I)(97), Discrete Mathematics(I)(98)
- **Homepage:** https://yxwang1215.github.io/

University of California, Los Angeles **Los Angeles, USA**
Visiting Student, Summer Session *Jun. 2024 – Oct. 2024*

- Mastered and applied mathematical modeling techniques via intensive study in Numerical Analysis.

Zhenhai High School of Ningbo **Ningbo, China**
Innovation Class (47 members), Jiaochuan Academy (Class 2) *Sep. 2020 – June 2023*

- Selected for the *Innovation Class* at Zhejiang's top-ranked high school; 15 classmates and 70 students school-wide are admitted to **Peking University** or **Tsinghua University** annually.

Publication

1. **Accelerating Diffusion Large Language Models with SlowFast: The Three Golden Principles**
   Q. Wei, Y. Zhang, Z. Liu, P. Zeng, **Y. Wang**, D. Liu, L. Zhang
   - **Accepted** to **ICLR 2026(CCF-A)**.
   - Proposed SlowFast Sampling, a dynamic strategy that adaptively alternates between exploratory and accelerated stages based on token certainty, convergence, and positional principles.
   - Achieved a 15.63x speedup on LLaDA and up to 34.22x when integrated with dLLM-Cache, outperforming LLaMA3 8B in throughput with minimal accuracy drop.
2. **UltraHiT: A Hierarchical Transformer Architecture for Generalizable Internal Carotid Artery Robotic Ultrasonography**
   T. Wang*, H. Jiang*, **Y. Wang***, Z. Sun, X. Yan, X. Li, G. Huang
   - **Co-First Author**; **Accepted** to **ICRA 2026 (CAA-A)**.
   - Proposed **UltraHiT**, a hierarchical Transformer for autonomous robotic ultrasonography, significantly improving generalization across diverse internal carotid artery (ICA) anatomies.
   - Engineered a multi-scale feature representation mechanism for precise probe navigation and robust vessel tracking, achieving state-of-the-art (SOTA) performance.

Project

1. **AudioKV: KV Cache Eviction in Efficient Large Audio Language Models**
   Y. Wang, P. He, X. Gui, X. Liu, J. He, X. Liu, X. Hu, L. Zhang
   - **First Author**; Submitted to **ICML 2026 (CCF-A)**.
   - Proposed **AudioKV**, an audio-aware KV cache compression framework for LALMs that prioritizes audio-critical attention heads via semantic-acoustic alignment.
   - **Spectral Score Smoothing (SSS)** enabling FFT-based filtering, which preserves acoustic temporal continuity while reducing memory overhead by 60

2. **Thinking inside the Mask: In-place Prompting in Diffusion LLMs**
   X. Jin, **Y. Wang**, Y. Gao, Z. Wen, B. Qi, D. Liu, L. Zhang
   - **Second Author**; Submitted to **ACL 2026 (CCF-A)**.
   - Proposed **In-Place Prompting**, a novel paradigm that integrates reasoning chains directly into the mask denoising process of Diffusion Large Language Models (dLLMs).
   - Developed a two-phase decoding strategy with early exiting, significantly enhancing inference efficiency and performance on complex reasoning benchmarks like GSM8K and MATH.
3. **Self Speculative Decoding for Diffusion Large Language Models**
   Y. Gao*, Z. Ji*, **Y. Wang**, B. Qi, H. Xu, L. Zhang
   - Submitted to **ACL 2026 (CCF-A)**
   - Developed **Self Speculative Decoding (SSD)** for Diffusion LLMs, accelerating inference via internal self-drafting and parallel verification without external draft models.
4. **AudioMarathon: A Comprehensive Benchmark for Long-Context Audio Understanding and Efficiency in Audio LLMs**
   P. He*, Z. Wen*, Y. Wang*, **Y. Wang**, X. Liu, J. Huang, Z. Lei, Z. Gu, X. Jin, J. Yang, *et al.*
   Submitted to **ACL 2026 (CCF-A)**.
5. **VA-Adapter: Adapting Ultrasound Foundation Model to Echocardiography Probe Guidance**
   T. Wang, H. Jiang, **Y. Wang**, Z. Sun, S. Song, G. Huang

Internships

LEAP Lab, Department of Automation, Tsinghua University
Research Assistant *Mar 2025 - September 2025*
- I participated in the implementation of the internal carotid artery ultrasound autonomous navigation project. Work was conducted onsite at Tsinghua University's Central Main Building. Our work has been accepted to **ICRA 2026**.

EPIC Lab, School of Artificial Intelligence, Shanghai Jiao Tong University
Research Assistant *August 2025 - January 2026 (Present)*
- Developed efficient algorithms for large language diffusion models, leveraging bidirectional attention as a high-performance alternative to autoregressive models.
- Spearheaded KV cache eviction research for Audio LLMs.

Selected Awards

Huawei Scholarship (less than 0.1%) *Dec. 2025*
Issued by Huawei Xi'an Research Institute
National Scholarship (Twice) *2023 -2024, 2024 - 2025*
National First Prize, National English Competition (NECCS) *Apr. 2025*
First Prize, 16th National Undergraduate Mathematics Competition *Nov. 2024*
First Prize, National Undergraduate Mathematical Modeling Contest *Dec. 2024*
National First Prize, *Vocabulary Star* National English Vocabulary Competition *Jan. 2024*
Honorable Mention, Mathematical Contest in Modeling (MCM) *Jan. 2024*

Activities

Xidian Inspur Club
President *2025 - 2026*
- **Club & Lab Management:** Directed daily operations and spearheaded club recruitment, while organizing academic workshops and orientations to grow the research community.