

Research Overview

My research centers on video understanding and generative models, with key areas of focus including:

- **Efficient Long-Sequence Modeling**, especially for long video inputs, using techniques like hybrid memory, token compression, RNNs, sparse attention, and linear attention mechanism.
- **Applications of Generative Models**, with an emphasis on techniques like masked image modeling for text-to-image synthesis, and a strong focus on enhancing efficiency in data usage and training.
- **Benchmarking and Evaluation**, creating complex and meaningful real-world challenges in video domains to probe the boundaries of model capabilities, while providing insights for future enhancement.

Education

M.S. CS 2023-2026	Zhejiang University Advisor: Gaoang Wang Rank: 1/87
Visiting Spring/Summer 2025	University of California San Diego Advisor: Zhuowen Tu
B.S. 2019-2023	Dalian University of Technology Software Engineering Rank: 4.1 / 4.5

Employment

Research Intern 2023-2024	Microsoft Research Asia Working on Video Understanding Mentor: Xun Guo
-------------------------------------	--

Awards and Honors

KAUST Rising Stars in AI Symposium (Saudi Arabia)	Nov 2025
Outstanding Paper at ICCV 2025 Knowledge-Intensive MR Workshop	Oct 2025
Lambda AI Cloud Credits Grant Sponsorship	Sept 2025
Graduate National Scholarship at Zhejiang University (2.4%)	Sept 2025
Graduate National Scholarship at Zhejiang University (2.4%)	Sept 2024
Undergraduate National Scholarship at Dalian University of Technology (2.3%)	Oct 2021

Selected Publications

The * sign denotes equal contribution.

Peer-Reviewed Papers

C5 Enxin Song*, Wenhao Chai*, Shusheng Yang, Ethan J. Armand, Xiaojun Shan, Haiyang Xu, Jianwen Xie, and Zhuowen Tu. **VideoNSA: Native Sparse Attention Scales Video Understanding**. In *ICLR*. 2026.

J1 Enxin Song*, Wenhao Chai*, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. **MovieChat+: Question-aware Sparse Memory for Long Video Question Answering**. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. 2025.

C4 Weili Xu*, Enxin Song*, Wenhao Chai*, Tian Ye, and Gaoang Wang. **Bringing RNNs Back to Efficient Open-Ended Video Understanding**. In *International Conference on Computer Vision (ICCV)*. 2025.

C3 Wenhao Chai*, Enxin Song*, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D. Manning. **AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark**. In *International Conference on Learning Representations (ICLR)*. 2025.

C2 Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Qing-Guo Chen, Xiangtai Li, Zhen Dong, Lei Zhu, Shuicheng Yan. **Meismonic: Revitalizing Masked Generative Transformers for Efficient High-Resolution Text-to-Image Synthesis**. In *International Conference on Learning Representations (ICLR)*. 2025.

C1 Enxin Song*, Wenhao Chai*, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. **MovieChat: From Dense Token to Sparse Memory for Long Video Understanding**. In *Computer Vision and Pattern Recognition (CVPR)*. 2024.

Workshop and Technical Reports

W3 Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. **Video-MMLU: A Massive Multi-Discipline Lecture Understanding Benchmark**. In *International Conference on Computer Vision (ICCV) Findings*. 2025.

W2 Weili Xu*, Enxin Song*, Wenhao Chai*, Tian Ye, and Gaoang Wang. **Bringing RNNs Back to Efficient Open-Ended Video Understanding**. In *Computer Vision and Pattern Recognition (CVPR) Workshop @ Efficient Large Vision Models*. 2025.

W1 Yichen Xu, Zihan Xu, Wenhao Chai, Zhonghan Zhao, Enxin Song, and Gaoang Wang. **Devil in the Number: Towards Robust Multi-modality Data Filter**. In *International Conference on Computer Vision (ICCV) Workshop @ DataComp*. 2023.

Preprints

P2 Ruizhe Chen, Zhiting Fan, Yang Shi, Enxin Song, Wenhao Chai, Tongkun Guan, Songtao Jiang, Ruilin Luo, Yuanxing Zhang, Zhibo Yang, Sibo Song, Shuai Bai, Junyang Lin, and Zuozhu Liu. **Can “Think with Videos” Boost Video Perception? A Benchmark for Fine-grained Video Understanding**. In *Under Review*. 2025.

P3 Enxin Song, Wenhao Chai, Xun Guo, Gaoang Wang, Jenq-Neng Hwang, Yan Lu. **Fantasy: Transformer Meets Transformer in Text-to-Image Generation**. In *OpenReview*. 2024.

Invited Talks

From Seeing to Thinking
Lambda AI

Virtual
Sept 2025

Video-MMLU: A Massive Multi-Discipline Lecture Understanding Benchmark
Workshop on Knowledge MR at ICCV 2025

Oahu, HI
Oct 2025

Teaching

ECE 445 Senior Design (Undergraduate)

Teaching Assistant, Zhejiang University - University of Illinois Urbana-Champaign

Spring 2024

Professional Service

Workshop Organization

Multimodal Video Agent

Workshop on Computer Vision and Pattern Recognition (CVPR) 2025

Nashville, TN

June 2025

Long-form Video Understanding Towards Multimodal AI Assistant and Copilot

Workshop on Computer Vision and Pattern Recognition (CVPR) 2024

Seattle, WA

June 2024

Conference and Journal Refereeing

IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)

2025

Neural Information Processing Systems (NeurIPS)

2025

International Conference in Learning Representations (ICLR)

2025, 2026

Computer Vision and Pattern Recognition (CVPR)

2025, 2026

Pattern Recognition and Computer Vision (PRCV)

2023, 2025

IEEE Transactions on Multimedia (TMM)

2024