Kunyang Li

Efficient Training and Inference for Video Models, World Models, and Agent Systems

kunyang.li@ucf.edu Homepage Scholar LinkedIn

Education

University of Central Florida — Jan. 2024 – Present
Ph.D. in Computer Science — Orlando, FL
- Advisors: Mubarak Shah, Yuzhang Shang
- Research on efficient training and inference of generative models, focusing on video models, world models, and agent systems

Nanyang Technological University — Aug. 2022 – Dec. 2023
M.Sc. in Artificial Intelligence (GPA: 4.25/5.0) — Singapore
- Advisor: Shijian Lu

University of Electronic Science and Technology of China — Sept. 2018 – Jun. 2022
B.Eng. in Software Engineering (GPA: 3.86/4.0) — Chengdu, China
- Captain of the university debate team; multiple-time Best Debater award winner

Publications

- **Kunyang Li**, Mubarak Shah, Yuzhang Shang. "*Attend Locally, Remember Linearly: Linear Attention as Cross-Frame Memory for Autoregressive Video Diffusion*." **In *Neural Information Processing Systems (NeurIPS)*, 2026.** *[PDF]*
- **Kunyang Li**, Mubarak Shah, Yuzhang Shang. "*PackCache: A Training-Free Acceleration Method for Unified Autoregressive Video Generation via Compact KV-Cache*." *arXiv:2601.04359*, 2026. *[PDF]*
- Abdul Mohaimen Al Radi, **Kunyang Li**, Yuzhang Shang, Mubarak Shah, Yu Tian. "*Keepsake: Selective Spatial Memory for Long-Horizon Video Generation*." *arXiv:2610.06588*, 2026. *[PDF]*
- Abdul Mohaimen Al Radi, **Kunyang Li**, Yuzhang Shang, Mubarak Shah, Yu Tian. "*Aero-World: Action-Conditioned Aerial Video Generation from Inertial Controls*." *arXiv:2605.19728*, 2026. *[PDF]*
- **Kunyang Li**, Jeffrey A. Chan Santiago, Sarinda D. Samarasinghe, Gaowen Liu, Mubarak Shah. "*GVD: Guiding Video Diffusion Model for Scalable Video Distillation*." *arXiv:2507.22360*, 2025. *[PDF]*
- Ashmal Vayani, Dinura Dissanayake, *et al.* (including **Kunyang Li**). "*All Languages Matter: Evaluating LMMs on Culturally Diverse 100 Languages*." **In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.** *[PDF]*
- **Kunyang Li**, Hai Nguyen, Joshua Lowe, Chenguang Zhao, Peace C. Madueme, Mehdi Hedjazi Moghari, Mubarak Shah, Pegah Khosravi, Yuzhang Shang. "*CineMR: Tool-Integrated Vision-Language Reasoning for Quantitative Cardiac MRI Assessment*." *arXiv:2610.01166*, 2026. *[PDF]*
- **Kunyang Li** et al. "*Personalized Video Generation: Progress, Applications, and Challenges*." In preparation. *Project Page*

Industry Experience

Adobe Research, Adobe — May 2026 – Aug. 2026
Research Scientist — San Jose, CA
- Mentor: Kevin Duarte
- Extended ARL^2 (NeurIPS 2026) to native 1K and 4K video generation, developing an efficient autoregressive model for high-resolution training and inference
- Designed a multi-patch linear attention tailored to high resolution that preserves fine-grained spatial details, and replaced 75% of the attention layers with our design to reduce the quadratic attention cost

Research Experience

Center for Research in Computer Vision, UCF Jan. 2024 – Present
Research Assistant Orlando, FL

- **Efficient Video Generation** (2025–Present): developing KV-cache compression, linear-attention memory, and attention acceleration methods for autoregressive video diffusion, leading to ARL^2 (NeurIPS 2026) and PackCache
- **WRIVA Challenge Project** (2025–2026): reconstructed a 3D point cloud of a real-world site from 5,540 ground, drone, and security-camera images, with block-wise reconstruction and GPS-based alignment; explored diffusion priors for 3D consistency
- **Cisco Research Project** (2024–2025): explored video dataset distillation and proposed GVD, the first diffusion-based video distillation method, which jointly distills spatial and temporal features and retains 78.29% of full-data accuracy on MiniUCF using only 1.98% of the frames

Multimedia & Interactive Computing Lab, NTU Aug. 2022 – Dec. 2023
Student Researcher Singapore

- **M.Sc. thesis:** investigated generated images from GANs and diffusion models as synthetic training data to improve the accuracy of downstream detection and segmentation models
- Analyzed how synthetic data affects training stability, generalization, and the domain gap between generated and real images

Hardware & Embedded Systems Lab, NTU May 2023 – Aug. 2023
Research Assistant Singapore

- Accelerated ViT and Swin Transformer inference by deploying INT8 quantization on NVIDIA FasterTransformer
- Implemented a program-level INT8 quantization simulator and used it to optimize the inference efficiency of quantized models

Knowledge & Data Engineering Lab, UESTC Sept. 2020 – Jun. 2022
Undergraduate Researcher Chengdu, China

- **B.Eng. thesis:** developed a skeleton-based fall detection method using a hybrid CNN–LSTM architecture to model human body structure and temporal motion dynamics
- Built an elderly fall detection and management platform on top of the method and entered it in the university student innovation and entrepreneurship competition

Teaching Experience

University of Central Florida Fall 2024 – Spring 2026
Graduate Teaching Assistant Orlando, FL

- **COT 4210, Discrete Structures II** (Spring 2026): formal languages and computation theory, including regular languages and the Chomsky hierarchy
- **COP 3502C, Computer Science I** (Spring 2025): algorithms and programming in C; led weekly labs for 40–50 students
- **CAP 5636, Advanced Artificial Intelligence** (Fall 2024): search, probabilistic reasoning, Markov decision processes, and deep learning; taught two lectures on probability theory as groundwork for the course's AI topics