Yuanhao Zou

Orlando, FL | yuanhaoz@ucf.edu | 734-358-6400| LinkedIn | GitHub | Scholar

Education

University of Central Florida, Orlando, PhD. in Computer Science — Aug 2025 – Present

- GPA: 4.0/4.0
- **Advisor:** **Chen Chen**
- **Research Interest:** Video Understanding, Efficient Vision Language Model, Multimodal Learning.

University of Michigan, Ann Arbor, MS in Electrical and Computer Engineering — Aug 2023 – May 2025

- GPA: 4.0/4.0

Central South University, China, BE in Computer Science and Technology — Sep 2019 – Jun 2023

- GPA: 88/100

Main Publications

- [**NeurIPS 2026 E&D, First Author**] Conditional Multi-Event Temporal Grounding in Long-Form Video
- [**ECCV 2026, Second Author**] CLARITY: Medical World Model for Guiding Treatment Decisions by Simulating Context-Aware Disease Trajectories in Latent Space
- [**CVPR 2026 Workshop, First Author**] How Should Video LLMs Output Time? An Analysis of Efficient Temporal Grounding Paradigms
- [**ICLR 2026, First Author**] ***A.I.R.***: Enabling Adaptive, Iterative, and Reasoning-based Frame Selection For Video Question Answering
- [**CVPR 2025, First Author**] Alignment, Mining and Fusion: Representation Alignment with Hard Negative Mining and Selective Knowledge Fusion for Medical Visual Question Answering
- [**CVPR 2025 Workshop, First Author**] MVCM: Enhancing Multi-View and Cross-Modality Alignment for Medical Visual Question Answering and Medical Image-Text Retrieval
- [**Knowledge-Based Systems, First Author**] HFA-UNet: Hybrid and Full Attention UNet for Thyroid Nodule Segmentation

Internship

United Imaging Intelligence (Full-Time) — May. 2026 – Aug. 2026

- **Multi-Modality Video Reasoning Agent using Reinforcement Learning:** Training an RL-based agent that works on Multi-Modality Videos (Thermal, Depth, and Multiview) to solve complex questions that require joint reasoning ability.

Axon (Part-Time) — Oct. 2025 – Dec. 2025

- **Efficient Video Temporal Grounding:** Developed efficient temporal grounding for Video LLMs on edge devices like Axon AB4; a continuous temporal-decoding paradigm enables a **1.5B** model to surpass **7B** baselines (**57.1** vs. 27.5 mIoU on Charades-STA) with lower latency via non-autoregressive decoding. **Outcomes:** First-Author **CVPR 2026 Workshop** paper.

Research Experience

University of Central Florida, **Prof. Chen Chen** Feb. 2025 – Present

- **Video Understanding:** Frame Selection Model integrates fast lightweight models like CLIP with heavy vision language models, boosting accuracy by up to **+7.0%** (NExT-QA) and **+4.5%** (LongVideoBench) while analyzing only ∼32 frames on average vs. 128, cutting VLM inference time by ∼**74%** (162s → 42s).
- **Efficient Vision Language Model:** Developing efficient vision language models deployed on edge devices.
- **CoMET-Bench / CoMET-Agent:** Built **600**-video, **2,789**-query CoMET-Bench and Rejection-F1 for joint counting, grounding, and rejection evaluation. Designed a training-free graph agent, lifting GPT-5 **F1@0.5 from 10.1% to 16.2%**. Qualcomm AI Research collaboration; first-author **NeurIPS 2026 E&D** paper.

Stony Brook University, **Prof. Zhaozheng Yin** Feb 2024 – Nov 2024

- **Medical Vision-Language model:** (1) Addressed challenges in cross-modality understanding and the underutilization of multi-view images in medical radiology datasets, and applied to Medical VQA and Medical Image-Report Retrieval. (2) Developed a unified representation alignment approach with hard negative mining and selective knowledge fusion, improving overall accuracy to **80.5%** on VQA-RAD (**+5.3%** over BiomedCLIP), **86.7%** on SLAKE, and **65.3%** on PathVQA.
- **Outcomes:** First Author of two research papers to **CVPR 2025**.

University of Nottingham Ningbo China (UNNC), **Prof. Xiangjian He** Sep 2022 – Aug 2023

- **Medical Image Segmentation:** Developed a deep learning network integrating U-Net and Transformer for hybrid attention and multi-scale fusion modules to address challenges with limited samples and small objects in thyroid nodule ultrasound segmentation.
- **Outcomes:** First-Author paper in Journal **Knowledge-Based Systems**, achieving **85.2%** Dice / **74.6%** mIoU on DDTI (**+2.4%** / **+1.0%** over prior SOTA) with strong cross-dataset generalization across three public thyroid benchmarks.

Other Publications

- [**International Workshop on Advanced Imaging Technology (IWAIT) 2024**] Multimodality semisupervised learning for ophthalmic biomarkers detection.
- [**International Workshop on Advanced Imaging Technology (IWAIT) 2024**] Alignment, Mining and Fusion: Representation Alignment with Hard Negative Mining and Selective Knowledge Fusion for Medical Visual Question Answering.

Technologies

Languages: Python, C++

Deep Learning: PyTorch, PyTorch Lightning, HuggingFace Transformers, Accelerate, DeepSpeed, vLLM

Libraries: NumPy, Pandas, OpenCV

Tools & Infrastructure: Linux, Git, Docker, Slurm, Multi-GPU / Distributed Training, Weights & Biases