Zihao Sheng

1415 Engineering Dr, Madison, WI 53706

☎ 608-421-8887 ✉ zihao.sheng@wisc.edu in linkedin.com/in/zihao-sheng-3b1534261 ⌂ Personal Website

Education

University of Wisconsin-Madison **Madison, WI**

Ph.D., AI for Autonomous Driving Jan. 2023 – Present

- 2025 Google PhD Fellowship – University Nominee

Shanghai Jiao Tong University **Shanghai, China**

M.S. in Control Engineering (Electrical & Computer Engineering focus) Sep. 2019 – Mar. 2022

Xi'an Jiaotong University **Xi'an, China**

B.Eng. in Automation (Electrical & Computer Engineering focus) Aug. 2015 – Jun. 2019

Work Experience

Bosch Research North America & Bosch Center for Artificial Intelligence **Sunnyvale, CA**

Research Intern, Mentor: Dr. Xin Ye, Dr. Jingru Luo Nov. 2025 – Present

Research Topic: Vision-Language-Action (VLA) Models for Autonomous Driving

- Developing a VLA model that generates future control actions based on **visual perception** while following **textual navigation instructions**.
- Augmenting the VLA model to jointly predict future images and actions during training, providing **dense self-supervision** that helps the model learn **world dynamics** and produce more informed control actions.

Research Projects

Safe Autonomous Driving via Foundation Models, Diffusion Models, and RL **Madison, WI**

Research Assistant, University of Wisconsin-Madison, Advisor: Prof. Sikai Chen Jan. 2023 – Present

Traffic Accident Scene Understanding with Multimodal LLM – supported by NVIDIA Academic Grant Program

- Developed and fine-tuned a **7B MLLM** to support pixel-level segmentation, temporal grounding, region-based visual question answering, and accident video description.
- Curated a **220K multimodal QA dataset** with bounding boxes, segmentation masks, and temporal boundaries.
- Achieved **20–40% higher BLEU/ROUGE** and up to **+40 mIoU/AP** over state-of-the-art baselines.

World Model for Controllable Traffic Video Generation – supported by NVIDIA Academic Grant Program

- Built and trained a **3B diffusion-based world model** to generate future driving video frames conditioned on **ego trajectories and other agents' motions**, enabling predictive and safety-critical scenario generation.
- Designed a **bbox-guided loss and generation module**, allowing fine-graine editing of nearby agents' motions.
- Achieved **11% lower FVD** over state-of-the-art diffusion baselines.

Vision Language Models for Reinforcement Learning in Safe Driving

- Unified **VLMs with RL** to replace manual reward engineering and enable safer, generalizable driving policies.
- Leveraged **CLIP** to compute semantic rewards from image–text alignment, with vehicle states for stable training.
- Achieved **10.5% lower collisions**, **+104% route completion**, and strong generalization in CARLA.

Selected Publication [Google Scholar] (* indicates co-first author)

- **Sheng, Z.**, Huang, Z., ... & Chen, S. (2025). Talk2Traffic: Interactive and Editable Traffic Scenario Generation for Autonomous Driving with Multimodal Large Language Model. In: ***CVPR 2025*** *WDFM-AD*. (Project Page)
- **Sheng, Z.**, Huang, Z., ... & Chen, S. (2025). SafePLUG: Empowering Multimodal LLMs with Pixel-Level Insight and Temporal Grounding for Traffic Accident Understanding. (under review) (Project Page)
- **Sheng, Z.***, Huang, Z.*, ... & Chen, S. (2025). VLM-RL: A Unified Vision Language Models and Reinforcement Learning Framework for Safe Autonomous Driving. *Transportation Research Part C.* (Project Page)
- **Sheng, Z.**, Huang, Z., ... & Chen, S. (2025). CurricuVLM: Towards Safe Autonomous Driving via Personalized Safety-Critical Curriculum Learning with Vision-Language Models. (under review) (Project Page)

Technical Skills

Languages: Python, Java, JavaScript, C/C++, C#, Julia, MATLAB
Technologies/Tools: PyTorch, Transformers, LLMs/VLMs, Diffusers, Linux, Distributed Training, Git, Unity, CARLA